Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations

Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations

Michael M. Desai[1*], Aleksandra M. Walczak[2*], and Daniel S. Fisher[3]
[1] *Department of Organismic and Evolutionary Biology, Department of Physics, and FAS Center for Systems Biology, Harvard University*
[2] *CNRS-Laboratoire de Physique Théorique de l'École Normale Supérieure,*
[3] *Department of Applied Physics and Department of Bioengineering, Stanford University*
[*] *These authors contributed equally to this work*
(Dated: August 17, 2012)

Positive selection distorts the structure of genealogies and hence alters patterns of genetic variation within a population. Most analyses of these distortions focus on the signatures of hitchhiking due to hard or soft selective sweeps at a single genetic locus. However, in linked regions of rapidly adapting genomes, multiple beneficial mutations at different loci can segregate simultaneously within the population, an effect known as clonal interference. This leads to a subtle interplay between hitchhiking and interference effects, which leads to a unique signature of rapid adaptation on genetic variation both at the selected sites and at linked neutral loci. Here, we introduce an effective coalescent theory (a "fitness-class coalescent") that describes how positive selection at many perfectly linked sites alters the structure of genealogies. We use this theory to calculate several simple statistics describing genetic variation within a rapidly adapting population, and to implement efficient backwards-time coalescent simulations which can be used to predict how clonal interference alters the expected patterns of molecular evolution.

I. INTRODUCTION

Beneficial mutations drive long-term evolutionary adaptation, and despite their rarity they can dramatically alter the patterns of genetic diversity at linked sites. Extensive work has been devoted to characterizing these signatures in patterns of molecular evolution, and using them to infer which mutations have driven past adaptation.

When beneficial mutations are rare and selection is strong, adaptation progresses via a series of selective sweeps. A single new beneficial mutation occurs in a single genetic background, and increases rapidly in frequency towards fixation. This is known as a "hard" selective sweep, and it purges genetic variation at linked sites and shortens coalescence times near the selected locus [1]. Most statistical methods used to detect signals of adaptation in genomic scans are based on looking for signatures of these hard sweeps [2–6].

Hard selective sweeps are the primary mode of adaptation in small to moderate sized populations in which beneficial mutations are sufficiently rare. However, in larger populations where beneficial mutations occur more frequently, many different mutant lineages can segregate simultaneously in the population. If the loci involved are sufficiently distant that recombination occurs frequently enough between them, their fates are independent and adaptation will proceed via independent hard sweeps at each locus. However, in largely asexual organisms such as microbes and viruses, and on shorter distance scales within sexual genomes, selective sweeps at linked loci can overlap and interfere with one another. This is referred to as clonal interference, or Hill-Robertson interference in sexual organisms [7, 8]. These interference effects can dramatically change both the evolutionary dynamics of adaptation and the signatures of positive selection in patterns of molecular evolution. We illustrate them schematically in Fig. 1.

We and others have characterized the evolutionary dynamics by which a population accumulates beneficial mutations in the presence of clonal interference [7, 9–13]. Many recent experiments in a variety of different systems have confirmed that these interference effects are important in a wide range of laboratory populations of microbes and viruses [14–18]. These theoretical and experimental developments have recently been reviewed by Park *et al.* [19] and Sniegowski and Gerrish [20].

Although this earlier theoretical work has provided a detailed characterization of evolutionary dynamics in the presence of clonal interference, it does not make any predictions about the patterns of genetic variation within an adapting population. In this paper, we address this question of how clonal interference alters the structure of genealogies, and how this affects patterns of molecular evolution both at the sites underlying adaptation and at linked neutral sites. This has become particularly relevant in light of recent advances that now make it possible to sequence individuals and pooled population samples from microbial adaptation experiments [18, 21–23].

We note that much recent work in molecular evolution and statistical genetics has analyzed related scenarios where adaptation involves multiple mutations, motivated by recent theoretical work [24–26] and empirical data from *Drosophila* [27] and humans [28, 29] that suggests that simple hard sweeps may be rare. This includes most notably analysis of the effects of "soft sweeps," where recurrent beneficial mutations occur at a single locus, or selection acts on standing variation at this locus [30–32]. Soft sweeps drive multiple genetic backgrounds to moder-

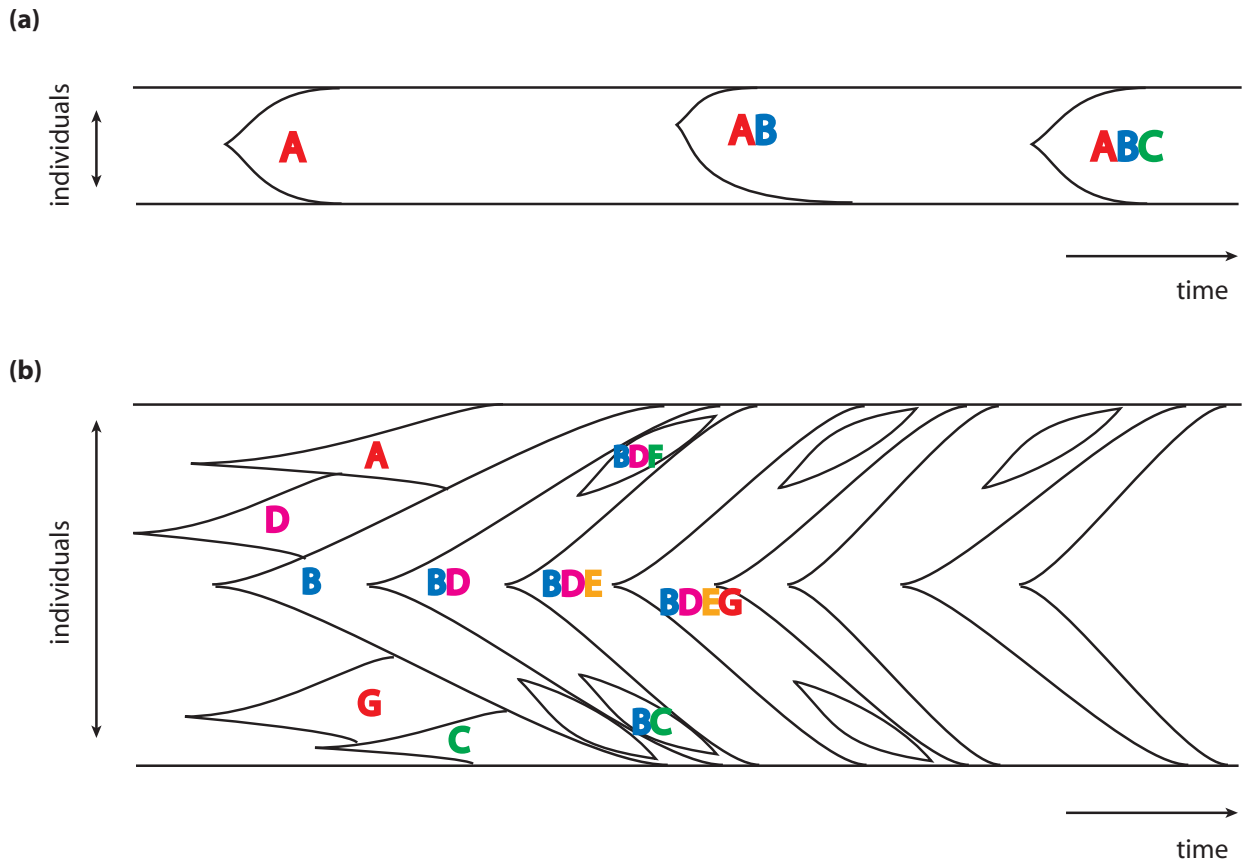

FIG. 1. Schematic of the evolutionary dynamics of adaptation. (**a**) A small population adapts via a sequence of selective sweeps. (**b**) In a large rapidly adapting population, multiple beneficial mutations segregate concurrently. Some of these mutant lineage interfere with each others' fixation, while others hitchhike together.

ate frequencies, leaving several deeper coalescence events and hence a weaker signature of reduced variation in the neighborhood of the selected locus than a hard sweep [33].

In contrast to the situation we analyze here, both hard and soft sweeps refer to the action of selection at a single locus. We consider instead a case more analogous to models in quantitative genetics, where selection acts on a large number of loci that all affect fitness. In other words, our analysis of clonal interference can be thought of as a description of polygenic adaptation, where selection favors the individuals who have beneficial alleles at multiple loci. Recent work has argued for the potential importance of polygenic adaptation from standing genetic variation [6, 34], loosely analogous to the case where soft sweeps act at many loci simultaneously [35, 36]. Our analysis in this paper, by contrast, describes polygenic adaptation via multiple new mutations of similar effect at many loci, where each locus has a low enough mutation rate that it would undergo a hard sweep in the absence of the other loci.

As with hard and soft sweeps, the signatures of this form of adaptation on nearby genomic regions are determined by how it alters the structure and timing of coalescence events. In this paper, we therefore focus on computing how clonal interference alters the structure of genealogies. This involves two basic effects. On the one hand, mutations at the many loci occur and segregate simultaneously, interfering with each others' fixation. This preserves some deeper coalescence events, as in a soft sweep. On the other hand, since the mutations occur at different sites, multiple beneficial mutations can also occur in the same genetic background and hitchhike together. This tends to shorten coalescence times, making the signature of adaptation somewhat more like a "hard sweep." Together, these effects lead to unique patterns of genetic diversity characteristic of clonal interference.

Our analysis of these effects is based on the fitness-class coalescent we previously used to describe the effects of purifying selection on the structure of genealogies [37]. This in turn is closely related to the structured coalescent model of Hudson and Kaplan [38]. We begin in the next section by describing our model, and summarize our earlier analysis of the rate and dynamics of adaptation

in the presence of clonal interference, which describes the distribution of fitnesses within the population [11]. We then show how one can trace the ancestry of individuals as they "move" between different fitness classes via mutations (our fitness-class coalescent approach). We compute the probability that any set of individuals coalesce when they are within the same fitness class. This leads to a description of the probability of any possible genealogical relationship between a sample of individuals from the population. Finally, we show how the distortions in genealogical structure caused by clonal interference alter the distributions of simple statistics describing genetic variation at the selected loci as well as linked neutral loci. We also use our approach to implement coalescent simulations analogous to those previously used to describe the action of purifying selection [39, 40], based on the structured coalescent method of Hudson and Kaplan [38]. These coalescent simulations can be used to analyze in detail how this form of selection alters the structure of genealogies.

Our results provide a theoretical framework for understanding the patterns of genetic diversity within rapidly evolving experimental microbial populations. Our analysis may also have relevance for understanding how pervasive positive selection alters patterns of molecular evolution more generally, but we emphasize that our work here focuses entirely on asexual populations or on diversity within a short genomic region that remains perfectly linked over the relevant timescales. In the opposite case of strong recombination, adaptation will progress via independent hard selective sweeps at each selected locus. Further work is required to understand the effects of intermediate levels of recombination, where the approach recently introduced by Neher *et al.* [41] may provide a useful starting point.

II. MODEL AND EVOLUTIONARY DYNAMICS

A. Model

We consider a finite haploid asexual population of constant size N, in which a large number of beneficial mutations are available, each of which increases fitness by the *same* amount s. We define U_b as the total mutation rate to these mutations. We neglect deleterious mutations and beneficial mutations with other selective advantages. We have previously shown that the dynamics in rapidly adapting populations are dominated by beneficial mutations of a specific fitness effect [11, 13, 42], so this model is a useful starting point, but we return to discuss these assumptions further in the Discussion. We also assume that there is no epistasis for fitness, so the fitness of an individual with k beneficial mutations is $w_k = (1+s)^k \approx 1 + sk$. This is the same model of adaptation we have previously considered [11] and is largely equivalent to models used in most related theoretical work on clonal interference [10, 19, 43]. We will later also consider linked neutral sites with total mutation rate U_n, but for now we focus on the structure of genealogies and neglect neutral mutations.

To analyze expected patterns of genetic variation, we must also make specific assumptions about how mutations occur at particular sites. We will consider a perfectly linked genomic region which has a total of B loci at which beneficial mutations can occur. We assume these mutations occur at rate μ per locus, for a total beneficial mutation rate $U_b = \mu B$. We will later take the infinite-sites limit, $B \to \infty$, while keeping the overall beneficial mutation rate U_b constant. Each mutation is assumed to confer the same fitness advantage s, where $s \ll 1$. We will also assume throughout that selection is strong compared to mutations, $s \gg U_b$, which allows us to use our earlier results in Desai and Fisher [11] as a basis for our analysis. Analysis of the opposite case where $s < U_b$ remains an important topic for future work, which could be based on alternative models of the dynamics such as the approach of Hallatschek [12]. Although our model is defined for haploids, our analysis also applies to diploid populations provided that there is no dominance (i.e., being homozygous for the beneficial mutation carries twice the fitness benefit as being heterozygous).

This model is the simplest framework that captures the effects of positive selection on a large number of independent loci of similar effect. However, the dynamics of adaptation in this model can be complex. Beginning from a population with no mutations at the selected loci, there is first a transient phase while variation at these loci initially increases. There is then a steady state phase during which the population continuously adapts towards higher fitness. Finally, adaptation will eventually slow down as the population approaches a well-adapted state. In this paper, we focus on the second phase of rapid and continuous adaptation, which has been the primary focus of previous work by us and others [11, 12, 19, 43]. Our goal is to understand how this continuous rapid adaptation alters the structure of genealogies and hence patterns of genetic variation. We begin in the next subsection by summarizing the relevant aspects of our earlier results for the distribution of fitness within the population.

B. The distribution of fitness within the population

In our model in which all beneficial mutations confer the same advantage, s, the distribution of fitnesses within the population can be characterized by the fraction of the population, ϕ_k, that has k beneficial mutations more or less than the population average. We refer to this as "fitness class k."

When N and U_b are small, it is unlikely that a second beneficial mutation will occur while another is seg-

regating. Hence adaptation proceeds by a succession of selective sweeps. In this regime, beneficial mutations destined to survive drift arise at rate $NU_b s$ and then fix in $\frac{1}{s}\ln[Ns]$ generations. Thus adaptation will occur by successive sweeps provided that

$$NU_b \ll \frac{1}{\ln[Ns]}. \quad (1)$$

When this condition is met, the population is almost always clonal or nearly clonal except during brief periods while a selective sweep is occurring. Thus we will have $\phi_0 = 1$ and $\phi_k = 0$ for $k \neq 0$.

In larger populations, however, new mutations continuously arise before the older mutants fix. Thus the population maintains some variation in fitness even while it adapts. The distribution of fitnesses within the population is determined by the balance between two effects. On the one hand, new mutants arise at the high-fitness "nose" of this distribution, generating new mutants more fit than any other individuals in the population. This increases the variation in fitness in the population. (While new mutations occur throughout the fitness distribution, the mutations essential to maintain variation are those that arise at the nose and generate new most-fit individuals.) On the other hand, selection destroys less-fit variants, increasing the mean fitness and decreasing the variation in fitness within the population. This is illustrated in Fig. 2.

We showed in previous work that this balance between mutation and selection leads to a constant steady state distribution of fitnesses within the population, measured relative to the current (and constantly increasing) mean fitness [11]. In this steady state distribution, the fraction of individuals with k beneficial mutations relative to the current mean in the population is typically

$$\phi_k = \phi_{-k} = Ce^{-\sum_{i=1}^{k} is\bar{\tau}}, \quad (2)$$

where $\bar{\tau}$ is defined below and C is an overall normalization constant that will not matter for our purposes. Note that the distribution ϕ_k is approximately Gaussian.

This distribution ϕ_k is cut off above some finite maximum k which corresponds to the nose of the distribution, the most-fit class of individuals. We define the *lead* of the fitness distribution, qs, as the difference between the mean fitness and the fitness of these most-fit individuals (so q is the maximum value of k; the most-fit individuals have q more beneficial mutations than the average individual). In Desai and Fisher [11], we showed that

$$q = \frac{2\ln[Ns]}{\ln[s/U_b]} \approx e^{-(s\bar{\tau}k)^2/2}. \quad (3)$$

This is illustrated in Fig. 2.

Above we have implicitly defined $\bar{\tau}$ to be the "establishment time," the average time it takes for new mutations to establish a new class at the nose of the distribution,

$$\bar{\tau} = \frac{\ln^2[s/U_b]}{2s\ln[Ns]}. \quad (4)$$

As we will see below, the characteristic time scale for coalescent properties will turn out to be the time for the fitness class at the nose to become the dominant population — i.e. for the mean fitness to increase by the lead of the fitness distribution. This takes q establishment times, so that the this "nose-to-mean" time is

$$\tau_{nm} \approx q\bar{\tau} \approx \frac{\ln(s/U_b)}{s}, \quad (5)$$

which is roughly independent of the population size for sufficiently large N. We note that no single mutant sweeps to fixation in this time: rather, a whole set of mutants comprising a new fitness class at the nose will come to dominate the population a time τ_{nm} later.

III. THE FITNESS-CLASS COALESCENT APPROACH

We now wish to understand the patterns of genetic variation within a rapidly adapting population in the clonal interference regime. To do so, we will use a fitness-class coalescent method in which we trace how sampled individuals descended from individuals in less-fit classes, moving between classes by mutation events. In each fitness class, there will be some probability of coalescence events. To calculate these coalescence probabilities, we must first understand the clonal structure within each fitness class: this we now consider.

A. Clonal structure

Each fitness class is first created when a new beneficial mutation occurs in the current most-fit class, creating a new most-fit class at the nose of the fitness distribution (see inset of Fig. 2). This new clonal mutant lineage fluctuates in size due to the effects of genetic drift and selection before it eventually either goes extinct or establishes (i.e. reaches a large enough size that drift becomes negligible). After establishing, the lineage begins to grow almost deterministically. Concurrently additional mutations occur at the nose of the distribution, also founding new mutant lineages within this most-fit class. This process is illustrated in Fig. 3a.

We wish to understand the frequency distribution of these new clonal lineages, each founded by a different beneficial mutation. In our infinite-sites model, each such lineage is genetically unique. We can gain an intuitive

understanding of this frequency distribution with a simple heuristic argument. After it establishes, the size of

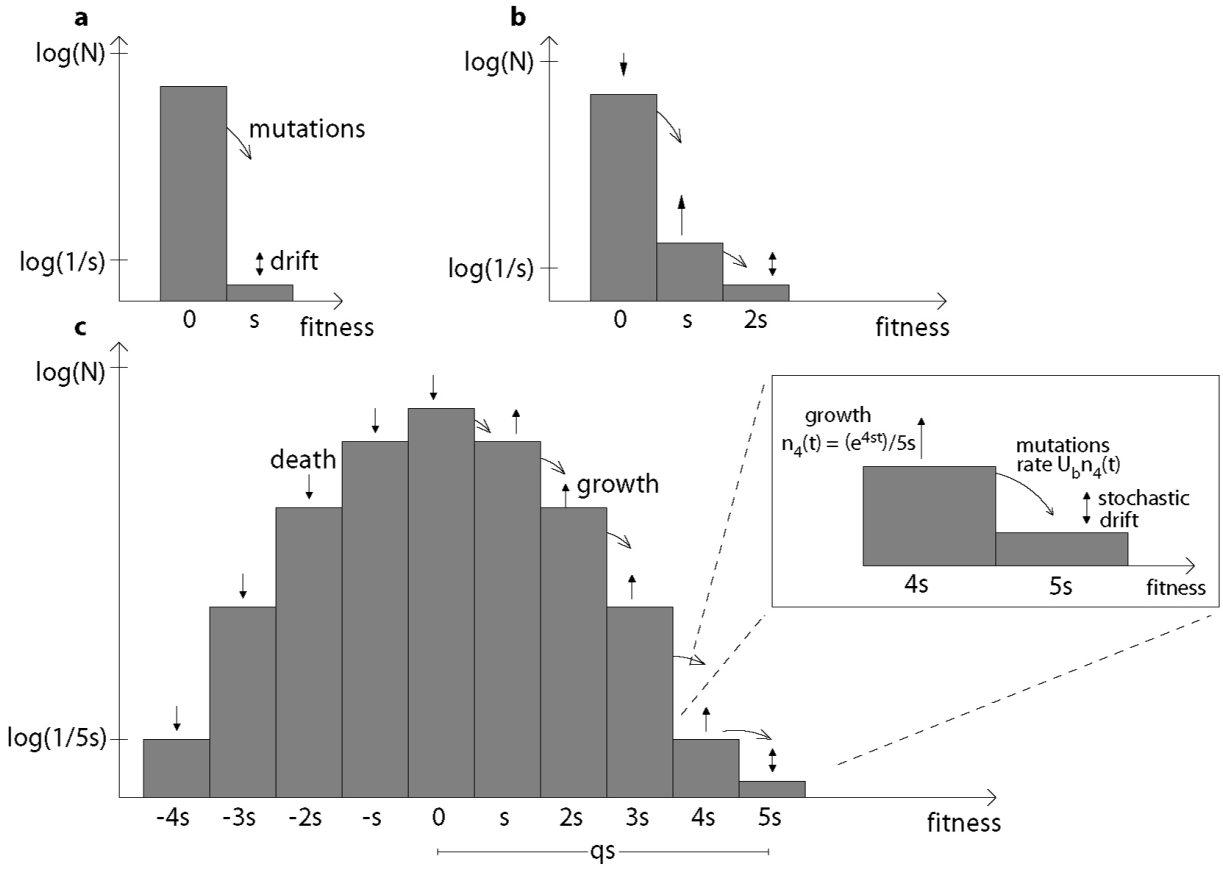

FIG. 2. Schematic of the evolution of large asexual populations, from Desai and Fisher [11]. The fitness distribution within a population is shown on a logarithmic scale. (**a**) The population is initially clonal. Beneficial mutations of effect s create a subpopulation at fitness s, which drifts randomly until it reaches a size of order $\frac{1}{s}$, after which it behaves deterministically. (**b**) This subpopulation generates mutations at fitness $2s$. Meanwhile, the mean fitness of the population increases, so the initial clone begins to decline. (**c**) A steady state is established. In the time it takes for new mutations to arise, the less fit clones die out and the population moves rightward while maintaining an approximately constant lead from peak to nose, qs (here $q = 5$). The inset shows the leading nose of the population.

the current most-fit class, $n_{q-1}(t)$, grows approximately deterministically according to the formula

$$n_{q-1}(t) = \frac{1}{qs} e^{(q-1)st}, \tag{6}$$

as we described in Desai and Fisher [11]. New mutations occur in individuals in this class at rate $U_b n_{q-1}(t)$, creating even more-fit individuals. Each new mutation has a probability qs of escaping genetic drift to form a new established mutant lineage. Thus the ℓ^{th} established mutant lineage at the nose will on average occur at roughly the time t_ℓ that satisfies

$$\int_0^{t_\ell} qsU_b n_{q-1}(t)dt = \ell. \tag{7}$$

Solving this for t_ℓ and then noting that the size, n_ℓ, of the ℓ^{th} established lineage will be proportional to $e^{qs(t-t_\ell)}$, we immediately find

$$\frac{n_\ell}{n_1} \approx \frac{1}{\ell^{1+1/q}}. \tag{8}$$

This provides a good estimate of the typical frequency distribution of clonal lineages within this fitness class at the nose, each lineage founded by a single new mutation.

The analysis above describes the clonal structure created as a new fitness class is formed, advancing the nose. After approximately $\bar{\tau}$ generations, the mean fitness of the population will have increased by s, and the growth rates of all the fitness classes we have described will decrease correspondingly. Thus we can strictly only use the calculations above up to some finite number of mutations, ℓ_{max}, after which all growth rates will have decreased due to the advance of the mean fitness of the population. Mutations will continue to occur after this time, but their frequency distribution will be slightly dif-

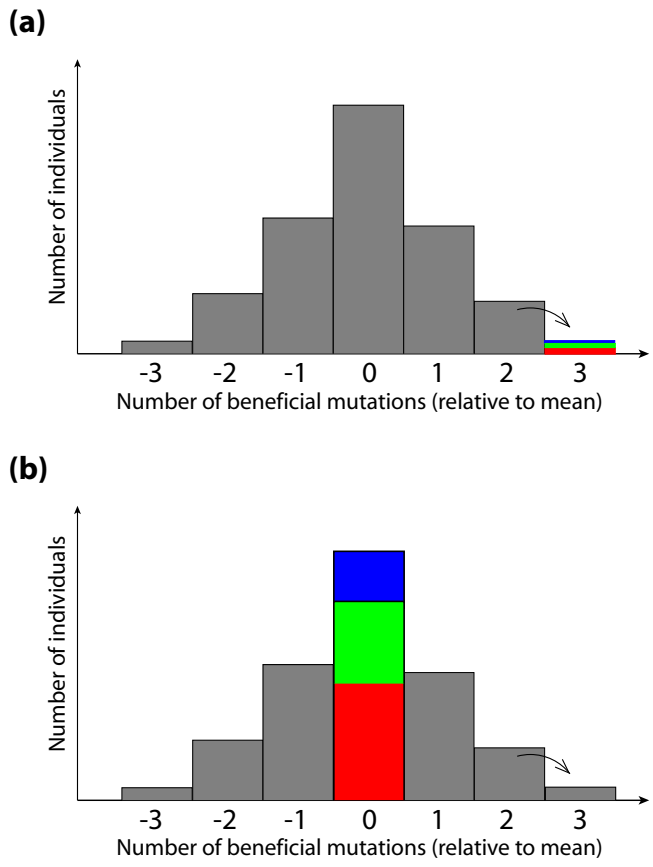

FIG. 3. Schematic of the establishment and fate of clonal lineages in a given fitness class, shown for a case where $q = 3$. (**a**) Three new clonal lineages (denoted in different colors) are established at the nose of the fitness distributions by three independent new mutations. These lineages have relative frequencies determined by the timing of these mutations. (**b**) After the population evolves for some time, the class that was at the nose of the distribution in (**a**) is now at the mean fitness. The class is still dominated by the three clonal lineages established while the class was at the nose (subsequent mutations represent only a small correction). These three clonal lineages have the same relative frequencies as when they were established at the nose; these relative frequencies remain "frozen" even as the population adapts.

ferent. Fortunately, in the strong selection regime we consider ($s \gg U_b$), the total contribution of all mutations after this point to the total size of the class is small compared to the contributions of the mutations that occur while this class is at the nose [11, 44]. We therefore neglect this cutoff to the number of mutations that occur at the nose, as well as the contribution of later mutations. This approximation will break down for very large samples. However, the errors it introduces can be shown to be relatively small even when considering quantities such as the time to the most recent common ancestor of the whole population.

Another important aspect of the dynamics that simplifies the behavior is that despite the changing growth rate of the fitness class as a whole, the *frequencies* of the established lineages within the class remain fixed. In other words, the clonal structure within the class remains "frozen" after it is initially created, rather than fluctuating with time (see Fig. 3b). As we will see, this and the neglect of late-arising mutations are good approximations in the regimes we consider here.

While our heuristic analysis provides a good picture of the typical frequency distribution of clonal lineages within each fitness class, it misses a crucial effect. Occasionally a new mutation at the nose will, by chance, occur anomalously early. This single mutant lineage can then dominate its fitness class. These events are quite rare, but when they do occur this single lineage can purge a substantial fraction of the total genetic diversity within the population. As we will see, these events together with less-rare but still early mutations are essential to the understanding the structure of genealogies within the population as they lead to a substantial probability of "multiple merger" coalescent events.

To capture these effects, we must carry out a more careful stochastic analysis of the clonal structure within each fitness class. As before, we focus on the clonal structure created when that class was at the nose of the fitness distribution, since it remains "frozen" thereafter. To do so, we note that the population size at the nose can be written as

$$n = \bar{n}(t) \sum_i \nu_i(t), \tag{9}$$

where $\bar{n}(t)$ reflects the average growth of all clones due to selection, and $\nu_i(t)$ reflects the stochastic effects of a clone generated from mutations at site i (of B total possible sites). At late enough times, the distribution of ν_i becomes time-independent, as shown previously [11]. This time-independent ν_i summarizes the combined effect of all the stochastic dynamics of mutations at this site that are relevant for the long-term dynamics. We showed that the generating function of ν_i is

$$G_i(z) = \left\langle e^{-z\nu_i} \right\rangle = \exp\left[-\frac{1}{B} z^{1-1/q}\right] = e^{-z^\alpha/B} = \left[1 - \frac{z^\alpha}{B}\right], \tag{10}$$

where angle brackets denote expectation values, the last equality follows for large B, and we have defined

$$\alpha \equiv 1 - \frac{1}{q}. \tag{11}$$

The total size of this fitness class is proportional to

$$\sigma \equiv \sum_{i=1}^{B} \nu_i. \tag{12}$$

This generating function $G_i(z)$ for the size of the clonal lineage founded at each possible site contains all of the

relevant information about the lineage frequency distribution, including the stochastic effects described above. Below we will use it to calculate coalescence probabilities within our fitness-class coalescent approach, which we now turn to.

B. Tracing Genealogies

To calculate the structure of genealogies, we take a fitness-class approach analogous to the one we used to analyze the case of purifying selection [37]. We first consider sampling several individuals from the population. These individuals come from some set of fitness classes with probabilities given by the frequencies of those fitness classes, ϕ_k. We note that in the purifying selection case, fluctuations in the ϕ_k due to genetic drift were a potential complication in determining these sampling probabilities. Here, these fluctuations are much less important provided that $U_b/s \ll 1$. We note however that fluctuations in different ϕ_k are correlated due to the stochasticity at the nose. Furthermore, averages of ϕ_k are far larger than their typical values due to rare fluctuations. Such fluctuations, which we discuss in detail elsewhere [45], may lead to some slight corrections to our results. But for most purposes, the typical values of the ϕ_k are what matters: thus we make the simple approximation that the probability of sampling one individual from class k_1 and a second from class k_2 is simply $\phi_{k_1}\phi_{k_2}$, with ϕ_k as given in Eq. (2). Analogous formulas apply for larger samples.

Each sampled individual comes from a specific fitness class k, and belongs to a specific clonal lineage within that class. This clonal lineage was created when this fitness class was at the nose of the distribution, approximately $(q-k)\bar{\tau}$ generations ago. It was created by a single new mutation in an individual from what is now fitness class $k-1$. That individual in turn belonged to some clonal lineage within class $k-1$, which in turn was created when that class was at the nose by a new mutation in an individual from what is now fitness class $k-2$, and so on.

We now describe the probability of a genealogy relating a sample of several individuals. Imagine, for simplicity, that we sampled two individuals that both happened to be in the same fitness class, k. If these individuals were from the same clonal lineage within that class, then they are genetically identical at all the B positively selected sites. We say they coalesced in class k and did so when this class was at the nose of the fitness distribution, approximately $(q-k)\bar{\tau}$ generations in the past. If these individuals were not from the same clonal lineage within the class, then they both descended from individuals, in what is now fitness class $k-1$, that got distinct beneficial mutations. If the individuals in which these mutations occurred are from the same clonal lineage within class $k-1$, we say the sampled individuals coalesced in class $k-1$. If so, they differ at two of the B positively selected sites, and coalesced when class $k-1$ was at the nose of the fitness distribution, approximately $[q-(k-1)]\bar{\tau}$ generations ago. If not, they descended from individuals, in what is now fitness class $k-2$, that got distinct beneficial mutations, and so on. We can apply similar logic to larger samples or when the individuals were sampled from different fitness classes. We illustrate this fitness-class coalescent process in Fig. 4.

We note that the probability a sample of individuals comes from the same clonal lineage is the same in each fitness class, since the clonal structure of the class was always determined when that class was at the nose of the distribution (nevertheless, conditional on some individuals coalescing in a class, the probability of additional coalescence events is substantially altered; see below). In addition, the coalescence probabilities do not depend on when the mutations occurred in the ancestral lineages of each sampled individual, since all clonal lineages were founded when a class was at the nose of the fitness distribution. These are major simplifications compared to the case of purifying selection, where the relative timings of mutations and the differences in clonal structure in different classes are important complications [37, 46].

To use the fitness-class coalescent approach to calculate the probability of a given genealogical relationship among a sample of individuals from the population, it only remains to calculate the probabilities that arbitrary subsets of these individuals coalesced within each fitness class. In the next section, we use the above described clonal structure to compute these fitness-class coalescence probabilities.

C. Fitness-class coalescence probabilities

We begin our calculation of the fitness-class coalescence probabilities by considering the probability that H individuals coalesce to 1 in a given class. We call this probability D_{H1}. This coalescence event will occur if and only if all H of these individuals are members of the same clonal lineage. The probability an individual is sampled from a clone of size ν is ν/σ, so summing over all possible clones we have

$$D_{H1} = \left\langle \sum_{i=1}^{B} \frac{\nu_i^H}{\sigma^H} \right\rangle \tag{13}$$

with $\sigma \equiv \sum_i \nu_i$. In Appendix A we use the expression for distribution of ν from Eq. (10), and take the $B \to \infty$ limit, to find

$$D_{H1} = \frac{\Gamma(H-\alpha)}{\Gamma(H)\Gamma(1-\alpha)}. \tag{14}$$

We can use a similar approach to calculate the probabilities of more complicated coalescence configurations.

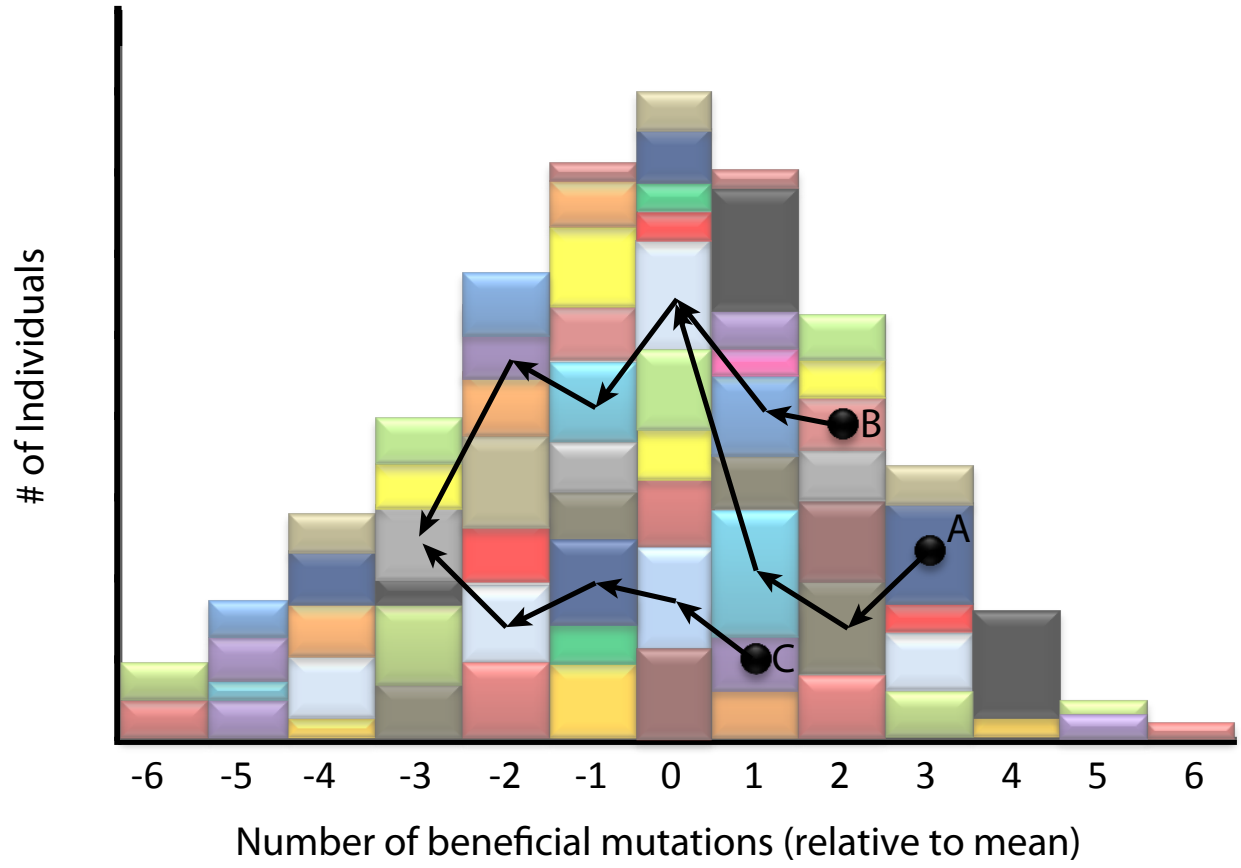

FIG. 4. Schematic of the fitness-class coalescent process. The distribution of fitnesses within the population is shown (here for a case where the nose is ahead of the mean by $q = 6$ beneficial mutations). Clonal lineages founded by individual beneficial mutations are shown in different colors within each fitness class. Three individuals (A, B, and C) were sampled from the population, from classes $k = 3$, $k = 2$, and $k = 1$ respectively. The ancestors of individuals A and B descended from individuals in the silver lineage in fitness class $k = 0$, and this individual shared a common ancestor with individual C in the gray lineage in class $k = -3$. Individuals A and B differ by 5 beneficial mutations, while individual C differs by 7 beneficial mutations from the common ancestor of B and C. Individuals A and B coalesce when the silver lineage in class $k = 0$ was originally created which occurred when this class was at the nose of the fitness distribution, $T_{AB} = 6\bar{\tau}$ generations ago. Individuals A, B, and C last shared a common ancestor when the gray lineage in class $k = -3$ was originally created when this class was at the nose of the fitness distribution, $T_{MRCA} = 9\bar{\tau}$ generations ago.

Consider the general situation where H individuals coalesce into K in a given fitness class, with h_1 individuals coalescing into lineage 1, h_2 individuals coalescing into lineage 2, and so on, up to h_K individuals coalescing into lineage K (note that $\sum_{j=1}^{K} h_j = H$). In Appendix A, we show that this probability, $C_{H,K,\{h_j\}}$, is given by

$$C_{H,K,\{h_j\}} = \frac{H\alpha^{K-1}}{K} \prod_{j=1}^{K} \frac{\Gamma(h_j - \alpha)}{\Gamma(h_j + 1)\Gamma(1 - \alpha)}. \qquad (15)$$

In order to compute any quantity that depends on genealogical topologies, it will be important to know not just that H individuals coalesced into K lineages, but that they did so in a specific configuration $\{h_j\}$. For example, if we have four individuals coalescing into two, this could occur by three of them coalescing into one and the other lineage not coalescing, or alternatively by two pairwise coalescence events. These different topologies will affect some aspects of molecular evolution such as the polymorphism frequency distribution. To compute these quantities, we must work with the full coalescence probabilities in Eq. (15).

However, the specific coalescence configurations do not affect non-topology-related quantities such as the total branch length, time to most recent ancestor, or any statistics that depend on these quantities (e.g. the total number of segregating sites S_n). To compute the statistics of these aspects of genealogies, we only need to know H and K. Thus it will be useful to sum the probabilities of all possible configurations $\{h_j\}$ that lead to a particular K. We call this total probability of H individuals

coalescing to K lineages D_{HK}. We have

$$D_{HK} = \sum_{\{h_j\}} C_{H,K,\{h_j\}}, \tag{16}$$

where the sum over the $\{h_j\}$ is constrained to values such that $\sum_{j=1}^{K} h_j = H$.

To compute D_{HK}, we first make the definition

$$f(H,K) = \sum_{\{h_j\}} \prod_{j=1}^{K} \frac{\Gamma(h_j - \alpha)}{\Gamma(h_j+1)\Gamma(1-\alpha)}, \tag{17}$$

and note that

$$D_{HK} = \frac{H}{K\alpha} f(H,K). \tag{18}$$

We can compute $f(H,K)$ using a simple contour integral,

$$f(H,K) = \frac{1}{2\pi i} \int \frac{dz}{z^{H+1}} \left[1-(1-z)^{\alpha}\right]^K, \tag{19}$$

where the integral is taken circling the origin. We can alternatively the generating function for $f(H,K)$,

$$R_f(z) \equiv \sum_{H=0}^{\infty} f(H,K) z^H. \tag{20}$$

In Appendix A, we show that

$$R_f(z) = \left[1-(1-z)^{\alpha}\right]^K. \tag{21}$$

We can now compute $f(H,K)$ for arbitrary H and K by noting that

$$f(H,K) = \frac{1}{H!} \frac{d^H}{dz^H} R_f(z)|_{z=0}, \tag{22}$$

and substitute this into Eq. (18) to compute D_{HK}. To give a few examples, we find

$$D_{21} = \frac{1}{q} \tag{23}$$

$$D_{31} = \frac{1}{2q}\left(1+\frac{1}{q}\right) \tag{24}$$

$$D_{32} = \frac{3}{2q}\left(1-\frac{1}{q}\right). \tag{25}$$

Taking more derivatives, we can easily make a table of $f(H,K)$ and evaluate any arbitrary D_{HK}. We note that in the large H limit, one can directly obtain $f(H,K)$ using saddle point evaluation of the contour integral defined above.

Note that the case of rapid adaptation, for which clonal interference is pervasive, corresponds to the case where q is reasonably large (conversely $q = 1$ corresponds to sequential selective sweeps, and our analysis does not apply in this limit). In the large-q regime, D_{21} is small. In neutral coalescent theory, the probability of a three-way coalescence event would then be even smaller: $D_{31} \sim D_{21}^2$. However, this is not the case here: the probability three lineages coalesce is of the same order as the probability two lineages coalesce, $D_{31} \sim D_{21}$, so "multiple-merger" coalescence events are not uncommon. This is a signature of the fact that occasionally a fitness class is dominated by a single large clone, as described above. When this happens, that clone dominates the structure of genealogies, as any ancestral lineages we trace through the fitness distribution are very likely to have originated from this single large lineage, and hence will coalesce within this fitness class. Although these anomalously large clones are rare, they are sufficiently common that they are responsible for a significant fraction of the total coalescence events, and they are responsible for tendency of genealogies to take on a more "star-like" shape.

IV. GENEALOGIES AND PATTERNS OF GENETIC VARIATION

From the results above for the probabilities of all possible coalescence events in each fitness class, we can calculate the probability of any genealogy relating an arbitrary set of sampled individuals. From these genealogies, we can in turn calculate the probability distribution of any statistic describing the expected patterns of genetic diversity in the sample.

We begin by neglecting neutral mutations and calculating the structure of genealogies in "fitness-class" space. That is, we consider individuals sampled from some set of fitness classes. We trace their ancestries backwards in time as they "advance" from one fitness class to the next, via mutational events, and calculate the probability that they coalesce in a particular set of earlier-established classes. Since each step in the fitness-class coalescent tree corresponds to a beneficial mutation, this immediately gives us the pattern of genetic diversity at the positively selected sites. We later consider how these "fitness-class" genealogies correspond to genealogies in real time, and use this to derive the expected patterns of diversity at linked neutral sites.

A. The distribution of heterozygosity at positively selected sites

We first describe the simplest possible case, a sample of two individuals. If we sample two individuals at random from the population, the first comes from class k_1 and the second from class k_2 with probability $\phi_{k_1}\phi_{k_2}$. If these two individuals coalesce in class ℓ, their total pairwise heterozygosity at positively selected sites, π_b, will be $(k_1-\ell)+(k_2-\ell) = k_1+k_2-2\ell$.

We can now calculate the average π_b given k_1 and k_2 by noting that

$$\left\langle \pi^b_{k_1,k_2} \right\rangle = |k_2 - k_1| + \left\langle \pi^b_{k,k} \right\rangle. \tag{26}$$

By conditioning on whether two individuals sampled from class k coalesce within that class (in which case they have $\pi_b = 0$), we have

$$\left\langle \pi^b_{k,k} \right\rangle = 0 D_{21} + (1 - D_{21}) \left[\left\langle \pi^b_{k,k} \right\rangle + 2 \right], \tag{27}$$

which implies

$$\left\langle \pi^b_{k,k} \right\rangle = \frac{2(1 - D_{21})}{D_{21}}. \tag{28}$$

Plugging this into the above, we find

$$\left\langle \pi^b_{k_1,k_2} \right\rangle = |k_2 - k_1| + \frac{2(1 - D_{21})}{D_{21}}. \tag{29}$$

We can now average this over k_1 and k_2 to find the overall average. Since k_1 and k_2 are approximately normally distributed with variance $1/(s\bar{\tau})$, their average absolute difference is $\sqrt{4/(s\bar{\tau}\pi)}$. Thus we have

$$\left\langle \pi^b_{k_1,k_2} \right\rangle = \sqrt{\frac{4}{\pi s \bar{\tau}}} + \frac{2(1 - D_{21})}{D_{21}}. \tag{30}$$

Note that for large q, the second term (corresponding to heterozygosity between individuals sampled from the same class) is approximately $2q$, while the first term is approximately $\sqrt{4q/\pi \log(s/U_b)}$, which is smaller by a factor of $1/\sqrt{2\pi \log(Ns)}$. This is because most individuals are much closer to the mean than to the nose, so that $|k_1 - k_2| \ll q$. In other words, a rough but very simple approximation is to assume that all individuals are sampled from the mean fitness class.

We can use a similar approach to compute the full probability distribution of π_b. We have

$$P(\pi^b_{k,k} = \gamma) = D_{21}\delta_{\gamma,0} + (1 - D_{21}) P(\pi^b_{k,k} = \gamma - 2), \tag{31}$$

which implies that

$$P(\pi^b_{k,k} = \gamma) = \begin{cases} D_{21}(1 - D_{21})^{\gamma/2} & \text{for } \gamma \text{ even} \\ 0 & \text{for } \gamma \text{ odd} \end{cases}. \tag{32}$$

We can then write the more general result

$$P(\pi^b_{k_1,k_2} = \gamma) = D_{21}\delta_{\gamma,k_1-k_2} + (1 - D_{21}) P(\pi^b_{k_1,k_2} = \gamma - 2), \tag{33}$$

from which we find

$$P(\pi^b_{k_1,k_2} = \gamma) = \begin{cases} D_{21}(1 - D_{21})^{\frac{\gamma-(k_1-k_2)}{2}} & \text{for } \frac{\gamma-(k_1-k_2)}{2} \text{ even and } \gamma \geq k_1 - k_2 \\ 0 & \text{otherwise} \end{cases}. \tag{34}$$

If desired, we can now average these results over the distributions of k_1 and k_2 to get the unconditional distribution of π_b. In Fig. 5a and Fig. 5b, we illustrate these theoretical predictions for the overall distribution of pairwise heterozygosity with the results of full forward-time Wright-Fisher simulations, for two representative parameter combinations. We see that the distribution of heterozygosity has a nonzero peak, and that the agreement with simulations is generally good.

We emphasize that our results for $P(\pi_b)$ describe the *ensemble* distribution of heterozygosity. That is, if we picked a single pair of individuals from each of many *independent populations*, this is the distribution of π_b one would expect to see. It is *not* the population distribution: if we were to pick many pairs of individuals from the same population, the π_b of these pairs would not be independent because much of the coalescence within individual populations occurs in rare classes that are dominated by a single lineage for which D_{21} is much higher than its average value. Thus if we measured the average π_b within each population by taking many samples from it, the distribution of this $\bar{\pi}_b$ across populations would be different from the distribution computed above. In order to understand these within-population correlations, we now consider the genealogies of larger samples.

B. Statistics in larger samples

We can compute the average and distribution of statistics describing larger samples in an analogous fashion to the pair samples. For example, consider the total number of segregating positively selected sites among a sample of 3 individuals, which we call S_{3b}. These three individuals are sampled (in order) from classes k_1, k_2, and k_3 respectively with probability $\phi_{k_1}\phi_{k_2}\phi_{k_3}$. For three individuals sampled from the same fitness class k, by conditioning on the coalescence possibilities within class k we find that the average total number of segregating posi-

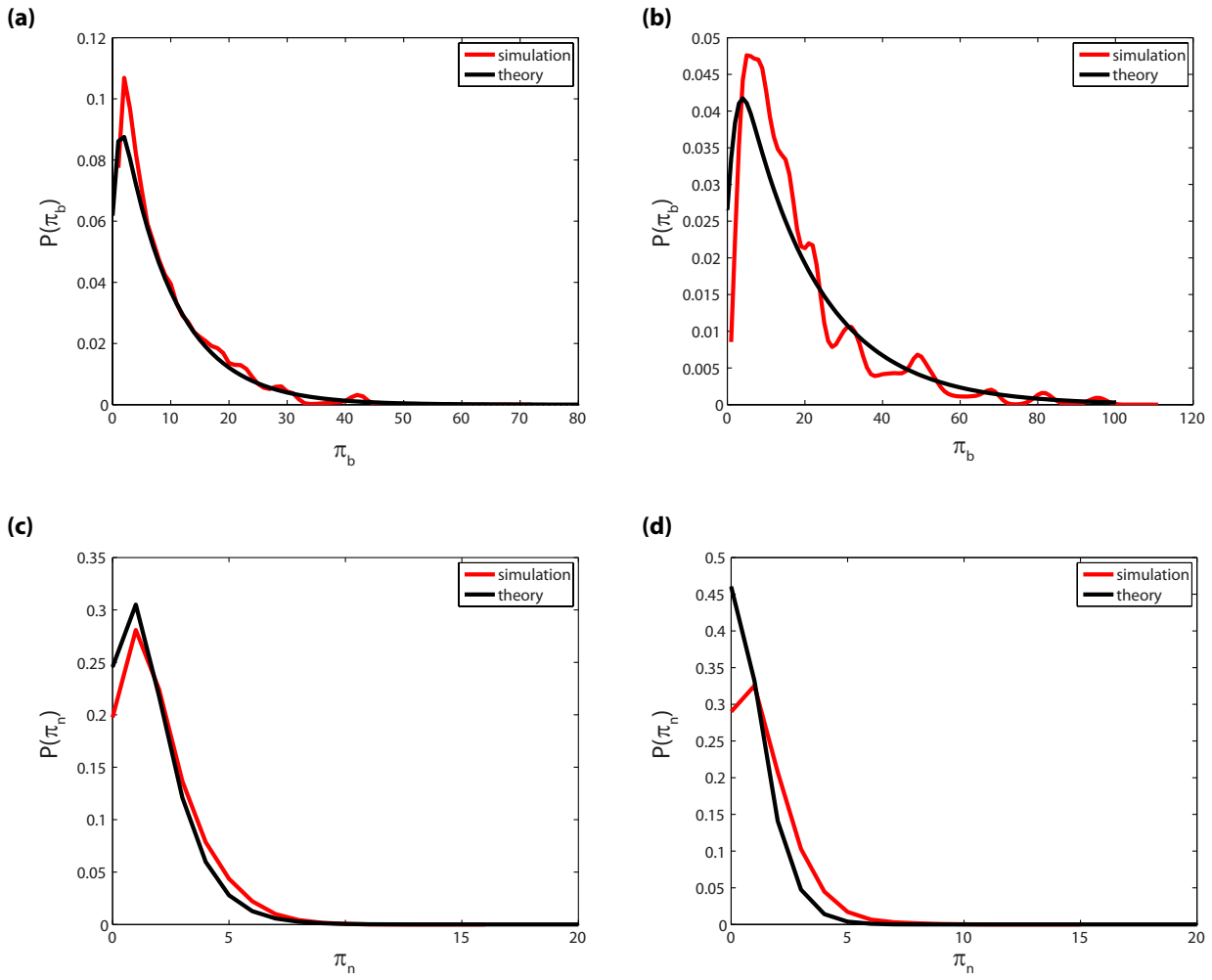

FIG. 5. The distribution of pairwise heterozygosity. (**a**) Comparison of our theoretical predictions for the distribution of pairwise heterozygosity at positively selected sites, π_b with the results of forward-time Wright-Fisher simulations, for $N = 10^7$, $s = 10^{-2}$, and $U_b = 10^{-4}$. Simulation results are an average over 56 independent runs, with 10^6 pairs of individuals sampled from each run. (**b**) Pairwise heterozygosity at positively selected sites for $N = 10^7$, $s = 10^{-2}$, and $U_b = 10^{-3}$. (**c**) Comparison of our theoretical predictions for the distribution of pairwise heterozygosity at linked neutral sites, π_n with the results of forward-time Wright-Fisher simulations, for $N = 10^7$, $s = 10^{-2}$, $U_b = 10^{-4}$, and $U_n = 10^{-3}$. (**d**) Pairwise heterozygosity at linked neutral for $N = 10^7$, $s = 10^{-2}$, $U_b = 10^{-3}$, and $U_n = 10^{-3}$.

tively selected sites is

$$\langle S_{kkk}\rangle = 0D_{31} + D_{32}\left[2 + \langle \pi^b_{k,k}\rangle\right] + D_{33}\left[3 + evS_{kkk}\right]. \tag{35}$$

Solving this for $\langle S_{kkk}\rangle$, we find

$$\langle S_{kkk}\rangle = \frac{2D_{32}/D_{21} + 3D_{33}}{D_{31} + D_{32}}. \tag{36}$$

More generally we have

$$\langle S_{k_1k_2k_2}\rangle = (1-D_{21})^{k_2-k_1}\left[2(k_2-k_1) + \langle S_{kkk}\rangle\right] + \sum_{i=0}^{k_2-k_1-1} D_{21}(1-D_{21})^i\left[k_2 - k_1 + \pi^b_{k,k} + i\right], \tag{37}$$

and even more generally we have

$$\langle S_{k_1k_2k_3}\rangle = k_3 - k_2 + \langle S_{k_1k_2k_2}\rangle. \tag{38}$$

If desired, we can average these over the distribution of k_1, k_2, and k_3 using the properties of differences of Gaussian random variables, as above. Alternatively, as in samples of size two, in large populations we can make the rough approximation that all sampled individuals come from the mean fitness class. Analogous calculations can be used to find the average number of segregating positively selected sites in still larger samples.

In Fig. 6 we illustrate some of these predictions (in practice samples are generated from coalescent simulations; see below) for samples of size 2, 3, and 10, and com-

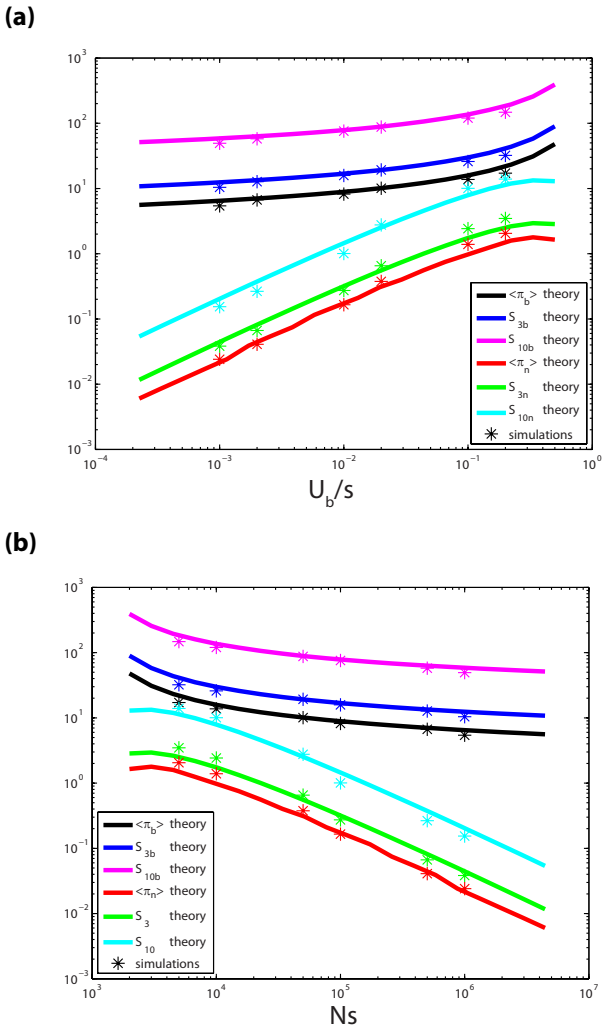

FIG. 6. Comparisons between theoretical predictions (from coalescent simulations) and forward-time Wright-Fisher simulations for the average pairwise heterozygosity and total number of segregating sites in samples of size 3 and 10 at positively selected sites and at linked neutral sites, (**a**) as a function of U_b/s and (**b**) as a function of Ns. In both panels, $N = 10^7$ and $U_b = 10^{-4}$ while s is varied. Forward-time Wright-Fisher simulation data represents an average over 56 forward simulation runs, with 10^6 pairs of individuals sampled from each run. Theoretical predictions generated using backwards-time coalescent simulations represent the average of 3×10^6 independently simulated pairs of individuals. Note that both (**a**) and (**b**) show the same data, plotted as a function of different parameters.

pare these to the results of forward-time Wright-Fisher simulations. We note that the agreement is generally good.

We can apply similar thinking to describe the distribution of the total number of segregating selected sites. First consider this distribution for a sample of size 3, all of which happen to be sampled from the same fitness class k, S_{kkk}. We have

$$P(S_{kkk} = \gamma) = D_{31}\delta_{\gamma,0} + D_{32}P(\pi^b_{k,k} = \gamma - 2) + D_{33}P(S_{kkk} = \gamma - 3). \quad (39)$$

We can multiply by z^γ and sum over γ to pass to generating functions, $U_3(z) \equiv \sum z^\gamma P(S_{kkk} = \gamma)$. This yields

$$U_3(z) = D_{31} + D_{32}z^2U_2(z) + D_{33}z^3U_3(z), \quad (40)$$

which we can solve to find

$$U_3(z) = \frac{D_{31} + z^2D_{32}U_2(z)}{1 - D_{33}z^3}, \quad (41)$$

where we have introduced the obvious notation.

More generally, we have that the total number of segregating sites among a sample of H individuals all chosen from the *same* fitness class k, which we will call S_H, has the distribution

$$P(S_H = \gamma) = D_{H1}\delta_{\gamma,0} + D_{H2}P(S_2 = \gamma - 2) + D_{H3}P(S_3 = \gamma - 3) + \ldots D_{HH}P(S_H = \gamma - H). \quad (42)$$

We can again pass to generating functions, giving

$$U_H(z) = D_{H1} + D_{H2}z^2U_2(z) + D_{H3}z^3U_3 + \ldots, \quad (43)$$

which we can easily solve to give

$$U_H(z) = \frac{D_{H1} + \sum_{\ell=2}^{H-1} z^\ell D_{H\ell}U_\ell(z)}{1 - D_{HH}z^H}. \quad (44)$$

It still remains to consider the distribution of the total number of segregating selected sites among H individuals chosen at random from arbitrary fitness classes. The general case becomes quite unwieldy to compute analytically, because we must average over all fitness classes in which internal coalescence events can occur. Computing these averages for the case of a sample of size three, we find that the generating function for the distribution of the total number of segregating positively selected sites among a sample of three individuals sampled from classes k_1, k_2, and k_3 is given by

$$W_3(z|k_1,k_2,k_3) = \frac{z^{k_1-k_3}U_2(z)D_{21}\left[1 - (zD_{22})^{k_1-k_2}\right]}{1 - D_{22}z} + D_{22}^{k_1-k_2}U_3(z). \quad (45)$$

Note that these distributions are all for samples each taken from an independently evolved population, rather than found from averaging many samples from each population and then finding the distribution of this across populations.

Analogous expressions can be computed for larger samples, but these involve ever more complex combinatorics. One may also wish to compute other statistics describing genetic variation in larger samples, such as the allele

frequency spectrum. While in principle it is possible to calculate analytic expressions for any such statistic using methods similar to those described above, in practice it is easier to use our fitness-class coalescent probabilities to implement coalescent simulations, and then use these simulations to compute any quantity of interest. We describe these coalescent simulations in a later section. Alternatively, for large populations we can make use of the rough approximation that all individuals are always sampled from the mean fitness class; we explore some consequences of this approximation further in a later section below.

C. Time in generations and neutral diversity

Thus far we have focused on the fitness-class structure of genealogies and the genetic variation at positively selected sites. We now describe the correspondence between our "fitness-class coalescent" genealogy and the genealogy as measured in actual generations. Fortunately, this correspondence is extremely simple: each clonal lineage was originally created by mutations when that fitness class was at the nose of the fitness distribution. Thus if we define the current mean fitness to be class $k = 0$, the current nose class will be at approximately $k = q$, and some arbitrary class k will have been created at the nose approximately $(q - k)\bar{\tau}$ generations ago. Although there is some variation in each establishment time, we neglect this variation throughout our analysis here, since it is small compared to the variation between coalescence times within clones in different classes. As we will see below, this approximation holds well in comparison to simulations in the parameter regimes we consider. This makes the correspondence between real times and step-times much simpler here than in our previous analysis of purifying selection, where the variation in real times, even given a specific fitness-class coalescent genealogy, was substantial [37].

The simple approximation of neglecting the variations in time of establishment of the fitness classes allows us to make a straightforward deterministic correspondence between the fitness-class coalescent genealogy and the coalescence times. We can then compute the expected patterns of genetic diversity at linked neutral sites: the number of neutral mutations on a genealogical branch of length T generations is Poisson distributed with mean $U_n T$. From this we can compute the distribution of statistics describing neutral variation (e.g. the neutral heterozygosity π_n or total number of neutral segregating sites in a sample S_n) from the corresponding statistics describing the variation at the positively selected sites. We illustrate these theoretical predictions for the distribution of neutral heterozygosity π_n in Fig. 5c and Fig. 5d, and compare these predictions to the results of full forward-time Wright-Fisher simulations. In Fig. 6 we also show our predictions (generated using the coalescent simulations described above) for the mean number of segregating neutral sites in samples of size 2, 3, and 10, compared to the results of forward-time Wright-Fisher simulations. We note that the agreement is good across the parameter regime we consider, though there are some systematic deviations for smaller values of U_b/s where our approximations are expected to be less accurate.

D. Time to the Most Recent Common Ancestor

Thus far we have considered the coalescence events at each mutational step separately: this is necessary to describe the full structure of genealogies. However, another important quantity of interest is the time to the most recent common ancestor — i.e. the coalescence time of the entire sample. We begin by considering this time measured in mutational steps, and then describe how this relates to the coalescence time measured in generations.

We can derive relatively simple expressions for the number of mutational steps to coalescence of an entire sample by directly calculating the probability of coalescence events over several steps at once. To do so, we note that since the dynamics at each mutational step are identical, the generating function of the number of individuals descended from a mutation at site i that occurred ℓ mutational steps ago, $\nu_i^{(\ell)}$, is given by

$$G_i^{(\ell)}(z) = \left\langle e^{-z\nu_i^{(\ell)}} \right\rangle = \exp\left[-\frac{1}{B} z^{\eta_\ell}\right], \tag{46}$$

where we have defined

$$\eta_\ell \equiv \alpha^\ell = (1 - 1/q)^\ell . \tag{47}$$

From this expression, we can immediately compute the distribution of the number of mutational steps to coalescence of H individuals sampled from the same fitness class, $J(H)$. The cumulative distribution of J is given by

$$F(H, \ell) \equiv \text{Prob}\left[J(H) \leq \ell\right] \approx \sum_{i=1}^{B} \left(\frac{\nu_i^{(\ell)}}{\sum_{i=1}^{B} \nu_i^{(\ell)}} \right)^H . \tag{48}$$

We can compute $\langle F(H, \ell) \rangle$ using identical methods to those used to calculate the fitness-class coalescence probabilities above, and find

$$\langle F(H, \ell) \rangle = \frac{\Gamma(H - \eta_\ell)}{\Gamma(H)\Gamma(1 - \eta_\ell)}. \tag{49}$$

From this, we find

$$\langle J(H) \rangle = \sum_{\ell=0}^{\infty} \left(1 - \langle F(H, \ell) \rangle\right). \tag{50}$$

Note that we could alternatively obtain expressions for $J(H)$ more directly from the fitness-class coalescence probabilities in a single step, by conditioning on the coalescence events that can happen in the first step in a similar way to that we used to compute $\langle \pi_b \rangle$ and $\langle S_{3b} \rangle$.

In the large-q limit, the ratios of these coalescence times (measured in mutational steps) in samples of different sizes are independent of q:

$$\frac{\langle J(3) \rangle}{\langle J(2) \rangle} = \frac{5}{4}, \quad \frac{\langle J(4) \rangle}{\langle J(2) \rangle} = \frac{25}{18}, \quad \frac{\langle J(5) \rangle}{\langle J(2) \rangle} = \frac{427}{288}. \tag{51}$$

These ratios are identical to those given by the Bolthausen-Sznitman coalescent [47], which has recently been shown to describe a number of other very different models of selection [48]. We return to this point in the Discussion. For large H we find

$$\frac{\langle J(H) \rangle}{\langle J(2) \rangle} \to \log\log H + \mathcal{O}(1). \tag{52}$$

These results suggest that there is a q-independent limiting process: we discuss this briefly below. We also note that the distribution of times to coalescence for large H is quite different than in the neutral case — the between-populations variation in $J(H)/\langle J(2) \rangle$ is only of order unity, compared to its mean of $\log\log H$. In contrast, for the neutral coalescent, the time to last common ancestor of the whole population has mean of $2\langle J(2) \rangle$ and random variations of the same order.

As with other aspects of genealogical structures, it is straightforward to convert these expressions for the coalescence times measured in mutational steps to the time in generations to the most recent common ancestor of a sample, $T_{MRCA}(H)$. Specifically, $J = \ell$ corresponds to the case where the most recent ancestor occurs ℓ mutational steps ago, so if the sampled individuals were from class k the time to the most recent common ancestor is $[q - (k - \ell)]\,\bar{\tau}$ generations. We note that for a sample of two this implies that the nose-to-mean time τ_{nm} is the characteristic time scale of the coalescent, as claimed above.

Thus far we have considered the most recent common ancestor of H individuals all sampled from the same fitness class k. However, in general we will typically sample individuals from a variety of different classes. In this case, we must sum over all possible internal coalescence events, until we reach a state where all remaining ancestral lineages are together in the same fitness class. This quickly becomes unwieldy in larger samples. In practice, it is easier to compute times to the most common recent ancestor in these cases using coalescent simulations based on our fitness-class coalescent approach, which we describe below.

As with other statistics described above, however, there is a simple approximation which is asymptotically correct for large populations: we can simply assume that all individuals are sampled from the mean fitness class. This approximation relies on the fact that most individuals sampled randomly from the population will have fitnesses close to the mean: within of order $\sqrt{v}$ of it. Thus the time differences between their establishments will typically be substantially smaller than the nose-to-mean time, τ_{nm}. As this is the time scale on which typical coalescent events take place, treating all the individuals as if they were in the dominant fitness class is a reasonable rough approximation. In this approximation, the results for the times to most common ancestor for samples of H can be simply obtained from the single-fitness class results above. We find:

$$\langle T_{MRCA}(2) \rangle \approx 2\tau_{nm}, \tag{53}$$

and in larger samples we have

$$\frac{\langle T_{MRCA}(3) \rangle}{\langle T_{MRCA}(2) \rangle} = \frac{9}{8}, \tag{54}$$

$$\frac{\langle T_{MRCA}(4) \rangle}{\langle T_{MRCA}(2) \rangle} = \frac{43}{36}, \tag{55}$$

$$\frac{\langle T_{MRCA}(5) \rangle}{\langle T_{MRCA}(2) \rangle} = \frac{715}{576}. \tag{56}$$

We note however that the dominant-fitness-class approximation is valid only in the limit that the lead of the population, qs, is much larger than the standard deviation of the fitness distribution, $\sqrt{v}$. As this ratio is $\sqrt{2\log(Ns)}$, in practice it never becomes very large.

E. The Frequency of Individual Mutations

An alternative way to compute many of the coalescent properties is to consider the fraction of the population with a particular mutation, which is closely related to the site frequency spectrum. The frequency of a given mutation at a particular site is determined by when that mutation occurred relative to others in its fitness class. In addition, its frequency at later times is determined by whether or not later mutations occur in its genetic background at each subsequent mutational step. Consider a mutation that occurred ℓ steps in the past, and define $f \equiv \frac{\nu}{\sigma}$ to be the fraction of the current nose class that its descendants constitute. The probability density of f is

$$\rho_\ell(f)df = \frac{df}{B} \frac{1}{\Gamma(\eta_\ell)\Gamma(1-\eta_\ell) f^{1+\eta_\ell}(1-f)^{1-\eta_\ell}}\ , \tag{57}$$

where as before we have defined $\eta_\ell = (1 - 1/q)^\ell$. Coalescent properties depend on averages of f^H. Summing over all B sites and using the standard integrals of powers of f and $1 - f$ expressed in terms of gamma functions,

we obtain immediately the result we had found above: $\langle F(H,\ell)\rangle = \frac{\Gamma(H-\eta_\ell)}{\Gamma(H)\Gamma(1-\eta_\ell)}$.

More generally, one can consider how the frequency of a mutation changes in time due to successive mutations in its lineage. If a given mutation has frequency g at one time, then a time $\ell\bar{\tau}$ later (after ℓ further beneficial mutations have occurred) the probability density of its frequency will be:

$$\rho_\ell(f|g)df = df\,\frac{\frac{g(1-g)}{\Gamma(\eta_\ell)\Gamma(1-\eta_\ell)f(1-f)}}{(1-g)^2\left(\frac{f}{1-f}\right)^{\eta_\ell} + g^2\left(\frac{1-f}{f}\right)^{\eta_\ell} + 2g(1-g)\cos(\pi\eta_\ell)}\ . \tag{58}$$

From this, quantities such as the variance of the probability of H individuals coalescing ℓ steps in the past and hence the variances in the coalescent times of H individuals can be computed.

In the limit of large q, the exponent η that parameterizes the time difference, $t = \ell\bar{\tau}$, is simply $\eta \approx e^{-t/\tau_{nm}}$. This is independent of q: only the "nose-to-mean" time that it takes for the new mutants to dominate the population matters. In this limit, a single mutational step occurs in a time that is a very small fraction, $\epsilon = 1/q$, of the nose-to-mean time τ_{nm}. The conditional probability of going from g to f in this step is

$$\rho_\ell(f|g)df \approx \frac{g(1-g)\epsilon\ df}{(f-g)^2 + \pi^2\epsilon^2[g(1-g)]^2}\ . \tag{59}$$

Eq. 59 is an approximate delta-function in $f-g$, as one would expect in the limit of a small time step. But it also corresponds to a probability per unit time of a jump from g to f of $\frac{1}{\tau_{nm}}dfg(1-g)/(f-g)^2$. Specifically it describes the genetic background either containing the mutation (frequency g) or not containing the mutation (frequency $1-g$) increasing in size by a factor between $1+h$ and $1+h+dh$ with rate $\frac{1}{\tau_{nm}}dh/h^2$ (with ϵ providing a small h cutoff). This corresponds to a continuous time birth process in a sub-population of (large) size n with rate per individual to give birth to k offspring, $\frac{1}{\tau_{nm}}\frac{1}{k^2}$. These considerations provide an alternative way to compute coalescent statistics.

F. Coalescent Simulations

We can use the fitness-class coalescence probabilities in Eq. (15) to implement an algorithm for coalescent simulations along the lines of Gordo *et al.* [39], using the structured coalescent framework of Hudson and Kaplan [38]. Specifically, to describe the diversity in a sample of n individuals, we first randomly sample their fitness classes independently from the distribution ϕ_k. We then start with the individual in the most-fit class, and trace back its ancestry as it steps through successive classes within the fitness distribution. When that individual enters a class with other individuals, we use Eq. (15) to determine the probabilities of all possible coalescence events in that class. We then continue to trace back the ancestry of the sample further through the distribution, allowing for coalescence events at each step according to the appropriate probabilities. We continue this procedure until all individuals have coalesced.

This simple coalescent algorithm produces a fitness-class coalescent tree drawn from the appropriate probability distribution of genealogies. We can then compute any statistic of interest describing this genealogy. By repeating this algorithm, we can obtain the probability distribution of the statistic. In practice this is a highly efficient procedure, since the coalescent simulations are extremely fast and the computational time required scales only with the size of the sample rather than the size of the population.

G. Comparison to Simulations

Our coalescent simulations represent an algorithmic implementation of our fitness-class coalescent, using all of the analytical expressions for the sampling and coalescence probabilities described above. Thus these coalescent simulations rely on all of the approximations underlying our method. To test the validity of these approximations and the accuracy of our fitness-class coalescent method, we compared the predictions of these coalescent simulations to full forward-time Wright-Fisher simulations of our model. These comparisons are illustrated in Fig. 5 and Fig. 6 and in Table I.

Our Wright-Fisher simulations were implemented assuming a population of constant size N, in which each generation consisted of a mutation and a selection step. In the mutation step, we independently choose the number of beneficial and neutral mutations within each extant genotype from the appropriate multinomial distribution. Each new mutation was assigned a unique index and all unique genotypes were tracked. In the selection step, we sample N individuals with replacement from the previous generation, using a multinomial sampling

u_b/s	Ns	D_{10} theory	D_{10} simulations
0.2000	5000	-3.3199	-3.3378
0.1000	10000	-3.3489	-3.3569
0.0200	50000	-3.3533	-3.3322
0.0100	100000	-3.3571	-3.4188
0.0020	500000	-3.3665	-3.3024
0.0010	1000000	-3.3717	-3.3670

TABLE I. Comparisons between theoretical predictions (from coalescent simulations) and forward-time Wright-Fisher simulations for Tajima's D [49] in a sample of size 10, D_{10}. Here $U_b = 10^{-4}$ and $N = 10^7$ while s is varied. Theoretical predictions are obtained by sampling 10^7 backward coalescent simulations. Forward-time simulation results are an average over 56 forward simulation runs, with 10^6 samples of $n = 2$ and $n = 10$ individuals.

weight adjusted for selective differences between individuals relative to the population mean fitness [50].

V. DISCUSSION

We have developed a fitness-class coalescent method to calculate how positive selection on many linked sites alters the structure of genealogies. This has allowed us to calculate how clonal interference shapes the patterns of genetic diversity in rapidly adapting populations. Our approach moves away from the traditional method of calculating the structure of genealogies in real time. Rather, we treat each mutational step from one fitness class to the next as an "effective generation," and trace how a sample of individuals descended by mutations through these fitness classes. In each "effective generation" we calculated the total probability of all possible coalescence events, Eq. (15). This allows us to calculate the structure of genealogies in this "fitness-class space," which directly corresponds to the genetic diversity at positively selected sites. We then converted this fitness-class coalescent to the genealogy in real time in order to calculate the expected patterns of neutral diversity.

We have shown that we can use this approach to compute analytic expressions for the distributions of several simple statistics describing patterns of molecular evolution. However, it is often easiest to compute expected patterns of variation using backwards-time coalescent simulations which explicitly implement the fitness-class coalescent algorithm using the distribution of the fraction of the population in each fitness class ϕ_k and the coalescence probabilities in Eq. (15) to simulate genealogies. These coalescent simulations are extremely efficient, and in practice it is usually faster to run millions of these backwards-time simulations than it is to numerically evaluate the sums over fitness classes involved in the corresponding exact analytic expressions. These coalescent simulations also have the advantage of being very similar in spirit to structured coalescent simulations that describe the effects of purifying selection (see e.g. Gordo *et al.* [39] and Seger *et al.* [40]), so they can in principle be used for parameter estimation and inference in analogous ways.

Our analysis throughout this paper is very similar in spirit to the fitness-class coalescent method we previously used to describe how purifying selection at many linked sites alters the structure of genealogies and patterns of molecular evolution [37, 46]. However, there are two important technical differences. First, in the case of purifying selection, fluctuations in the frequencies of each fitness class ϕ_k due to genetic drift can be substantial in certain parameter regimes. These fluctuations are particularly important near the nose of the distribution, where they can lead to effects such as Muller's ratchet. Although individuals are unlikely to be sampled from this nose, they are very likely to coalesce there. Neglecting these fluctuations was therefore an important approximation that substantially restricted the regime of validity of our analysis. By contrast, in the case of positive selection, fluctuations in the sizes of each fitness class are negligible (except at the nose) across a broad range of relevant parameter values. Furthermore, fluctuations at the nose are much less important for patterns of diversity than in the case of purifying selection, because individuals are unlikely to either be sampled there or to coalesce there. This reflects a fundamental difference between the neutral and purifying selection processes and the rapid adaptation dynamics analyzed here. For the former, genetic drift plays a key role in driving the fluctuations, while for the latter, genetic drift is almost irrelevant: the fluctuations are dominated by the stochasticity in the timings of the beneficial mutations that occur near the nose of the fitness distribution.

A second key simplification of our analysis of positive selection, compared to the purifying selection case, is that the clonal structure of each fitness class becomes effectively "frozen" once that class is no longer at the nose of the fitness distribution. This means that coalescence probabilities are identical in all fitness classes which stands in contrast to the case of purifying selection, where the clonal structure within all classes is constantly changing. This also avoids the need to carefully analyze the timing and order of mutation events in the history of a sample and simplifies the mapping between our fitness-class coalescent genealogy and the genealogy measured in real time.

Our results demonstrate how positive selection on many linked sites distorts the structure of genealogies away from neutral expectations. We show several examples of these selected genealogies, for various different parameter values, in Fig. 7. The most striking qualitative conclusion of our analysis is that multiple merger events, where several ancestral lineages coalesce into one

in a single effective generation, occur with comparable probabilities to pairwise coalescence events. We note that these events are multiple mergers within a single effective generation in our fitness-class coalescent, and hence are not actually multiple mergers within a single real generation. However, these events happen very close together in real time compared to the other relevant timescales, so they will appear as effectively instantaneous. This leads to a more "starlike" shape of genealogical trees. This signature is characteristic of the action of positive selection; our analysis here illustrates how starlike we expect genealogies to be (and how many deeper coalescence events are preserved) given the interplay between interference and hitchhiking effects characteristic of this rapid adaptation regime. It may prove useful in future work to analyze this specific situation in the context of more general models of the coalescent with multiple mergers [51].

We note that the characteristic time scale of the coalescence is the "nose-to-mean" time, τ_{nm}, which is the time after which the collection of new mutants at the nose take to dominate the population. In units of this time, trees for different values of q become statistically similar for large q. One striking feature, that occurs roughly once each τ_{nm}, is the coalescence of a substantial fraction of all the (remaining) lineages at a single time step: this is caused by one new beneficial mutation occurring so much earlier than typical that its descendants represent a substantial fraction of the population in the nose. Examples of this can be seen in Fig. 7. Another perhaps-surprising feature of the genealogies in large samples is that some aspects are *less* variable from one population to another than neutral coalescent trees, while other aspects are more variable. In the recent past, for times much shorter than the mean coalescence time of pairs of individuals, neutral coalescent trees, tend to be rather similar, while the multiple-coalescence events that characterize the positively selected genealogies cause larger variations between populations. In contrast, the time to last common ancestor of large samples is broadly distributed for neutral trees but narrowly distributed (at least asymptotically) for positively selected trees.

Because individuals are unlikely to be sampled from near the nose of the distribution, the initial coalescence events in the history of the sample are typically in the bulk of the fitness distribution. Since these coalescence events happened well in the past when these classes were at the nose of the distribution, the terminal branches in the genealogies of a sample are likely to be longer compared to internal branches than we would expect under neutrality. In other words, recent branches of genealogies are longer relative to more ancient branches. This effect is qualitatively similar to the situation in which effective population size declines as time recedes into the past: this has long been recognized as a general signature of the effects of both purifying and positive selection. It leads to an excess of singleton mutations in the site frequency spectrum, and the negative values of Tajima's D that we have observed. However, clonal interference mitigates these effects relative to a hard selective sweep.

Our results also demonstrate that even when beneficial mutations are rare compared to neutral mutations, $U_b \ll U_n$, positively selected sites can still contribute a significant fraction of the total genetic variation observed in a population. For example, in a sample of two individuals the total heterozygosity at positively selected sites will typically be several times q. The typical neutral heterozygosity, on the other hand, will be of order $\pi_n \sim U_n \tau_{nm}$. Thus even when $U_n \gg U_b$, π_b will often be comparable to or even greater than π_n. This is consistent with the general observation in microbial evolution experiments that a substantial fraction of observed mutations are beneficial [18, 21–23]. The fact that positively selected sites can be a significant fraction of the polymorphisms emphasizes the importance of understanding the patterns of diversity at these sites, which have distinct patterns compared to linked neutral variation and hence may provide important signatures in sequence data of adaptation that involves clonal interference.

Our predictions for the structure of the fitness-class genealogies depend on the population size, mutation rate, and strength of selection only through the combinations $\log[Ns]$ and $\log[U_b/s]$. The timescales in generations are also proportional to the inverse of the strength of selection. Thus the patterns of genetic variation in an adapting population depend only very weakly (logarithmically) on population size and mutation rate in the large-q regime where clonal interference is pervasive, suggesting that there is limited power to infer these parameters from patterns of molecular evolution. This is a consequence of the fact that the evolutionary dynamics are also only very weakly dependent on these parameters in the clonal interference regime.

We have seen that in the large-q limit of our model, the ratios of the number of mutational steps to the most recent common ancestors in samples of different sizes are exactly equivalent to those expected in the Bolthausen-Sznitman coalescent [47]. This is identical to the limiting behavior of these ratios in several very different models of selection recently studied by Brunet, Derrida, and others [52–56]; see [48] for a recent review. The reason for this equivalence between very different models remains unclear, but suggests a degree of universality: an interesting topic for future work. We emphasize, however, that the times to most recent ancestors in our model reduce to the Bolthausen-Sznitman ratios only when measured in mutational steps and only when all individuals are sampled from the same fitness class. The ratios of time to most recent common ancestors, measured in generations, have a different form. Nevertheless, in the limit of very large q, almost all the individuals will have fitness much closer to the mean than to the nose. As the rate of coalescence is proportional to the difference between the mean and

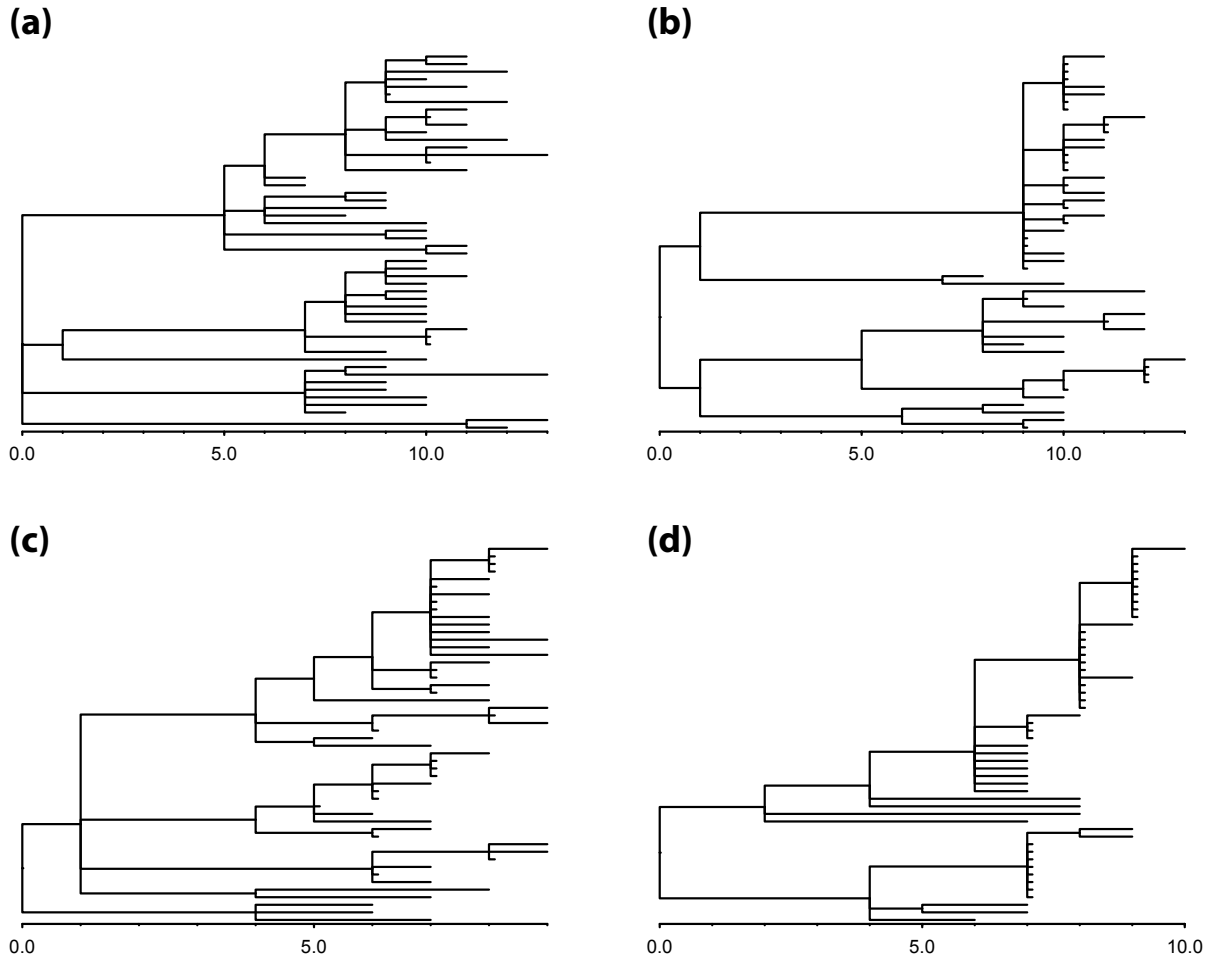

FIG. 7. Examples of fitness-class coalescent genealogies in samples of size 50 from forward-time Wright-Fisher simulations. The tips of each tree correspond to individuals sampled from the present. Each tip is placed horizontally according to the fitness class from which that individual was sampled (classes are numbered according to the number of beneficial mutations relative to the most recent common ancestor of the sample). Coalescence events are depicted according to the fitness class in which they occurred. Each unit of time on the horizontal axis corresponds to one beneficial mutation, so that two individuals separated by a branch length of ℓ have $\pi_b = \ell$. These fitness-class genealogies can be converted to genealogies in real time by using our approximation that all coalescent events happen when the relevant class was at the nose of the fitness distribution. Note that the characteristic time for coalescence is the time it takes for q successive beneficial mutations: this varies considerably with the parameters used. In all trees, $N = 10^7$ and $U_b = 10^{-4}$. (**a**) An example of a genealogical tree for $s = 10^{-3}$. (**b**) An example of a tree for $s = 5 \times 10^{-3}$. (**c**) An example of a tree for $s = 10^{-2}$. (**d**) An example of a tree for $s = 5 \times 10^{-2}$.

the nose, the approximation of sampling only from the largest fitness class is asymptotically good. The modifications of the Bolthausen-Sznitman ratios are then simply determined by adding the nose-to-mean time, (which turns out to be equal to the mean pairwise correlation time), to all the coalescent times.

Our analysis in this paper has focused on the simplest possible model of positive selection on a large number of linked sites, and we have neglected many potential complications. For example, we have assumed that epistatic interactions between mutations can be neglected, and that the total potential supply of beneficial mutations is not significantly depleted over the course of adaptation. This is consistent with our focus on rapidly adapting populations in the large-q clonal interference regime. As a population approaches a fitness peak, these approximations will likely fail and the dynamics of adaptation and patterns of genetic variation may either become more complex, or return to the regime where further adaptation is driven by isolated selective sweeps. We have also focused exclusively on beneficial mutations which all have the same fitness effect s, and have neglected both deleterious mutations and beneficial mutations which confer different fitness effects. This is justified by earlier work

by us and others that suggests that in rapidly adapting populations, clonal interference ensures that evolution is dominated by beneficial mutations that confer a specific fitness advantage [13, 42, 43]. However, we have recently analyzed the evolutionary dynamics within a population in a model which explicitly allows for a distribution of fitness effects of beneficial mutations [13]. We and others have also analyzed the case where a mix of both beneficial and deleterious mutations are possible [10, 43, 57]. Those works describe the variation in fitness within populations in these more complex models and hence could form the basis for a more complex version of the fitness-class coalescent method we have used here. This generalized fitness-class coalescent would admit the possibility of mutational steps of various different sizes and towards both lower and higher fitness.

An alternative approach by one of us allows for beneficial mutations to have a variety of different effects, without making reference to fitness classes [45]. As long as the distribution of fitness effects of potential beneficial mutations falls off faster than a simple exponential for large s, the dynamics in large populations is dominated by mutations with s close to some value, $\tilde{s}$ [13, 45]. In this case, most properties of the dynamics on time scales longer than the nose-to-mean time τ_{nm} are quite universal (and more strongly so when $v/\tilde{s}^2$ is large). As τ_{nm} is also the time scale of the coalescence, this suggests that the coalescent statistics should also be universal. The continuous time results quoted above for the evolution of the frequency of a sub-population emerges naturally in this more general analysis, and indeed correspond to the universal limit of asymptotically-large populations [45]. In the alternative regime where the distribution of fitness effects of potential beneficial mutations falls of more slowly than exponentially, mutations can jump from the bulk of the distribution to the lead. These play an important role in the dynamics, and cause q to remain small even for asymptotically large populations [11]. The behavior is then less universal, but this situation is likely to be relevant in real populations, especially in the initial stages of adaptation to a new environment. Further study into these effects of the distribution of effects of beneficial mutations, of initial transient dynamics, and of large numbers of deleterious mutations are interesting topics for future research.

The final simplification of our analysis is its focus on purely asexual populations: we have neglected the effects of recombination. Thus our results are primarily applicable to interpreting the patterns of genetic variation in asexual microbial evolution experiments, though they may also be relevant to sexual organisms on short genomic distance scales within which recombination is rare on the relevant timescales. We note however that our results provide an essential ingredient for predicting the effects of infrequent recombination on the evolutionary dynamics. Specifically, we can use our predictions for the genetic variation between a pair of individuals sampled from the population to predict the distribution of fitnesses of recombinant offspring resulting from sex between these individuals. This in turn determines how rare recombination alters the evolutionary dynamics and the distribution of fitnesses within the population. It may prove possible to then in turn calculate how these shifts in evolutionary dynamics alter the patterns of genetic diversity in the population. These extensions of our approach to analyze the effects of recombination on both evolutionary dynamics and patterns of molecular evolution are an important direction for future research.

VI. ACKNOWLEDGMENTS

We thank Richard Neher, Boris Shraiman, Thierry Mora, Lauren Nicolaisen, Benjamin Good, Elizabeth Jerison, and John Wakeley for many useful discussions. MMD acknowledges support from the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, and the Harvard Milton Fund. DSF acknowledges support from the National Science Foundation via DMS-1120699.

APPENDIX A: COALESCENCE PROBABILITIES

In this Appendix, we carry out the calculations of coalescence probabilities in detail. Consider H individuals who coalesce into K lineages, with h_1 individuals coalescing into lineage 1, h_2 individuals coalescing into lineage 2, and so on, up to h_K individuals coalescing into lineage K. We note that $\sum_{j=1}^{K} h_j = H$. We begin by asking the probability that H individuals coalesce into K lineages at a *specific* set of K sites (out of the total of B) in the genome: call these sites 1 through K in the genome, for concreteness. We also assume for now that the H individuals coalesce in a *specific* way into these K lineages: i.e. individual 3 coalesces into the lineage at site 5, etc). We denote the frequency of the lineage at site j in the genome by f_j; so that $f_j = \frac{\nu_j}{\sigma}$. We denote by A the probability that the H individuals coalesce into the K lineages at these specific sites according to the specific configuration $\{h_j\}$.

Given these definitions, we have:

$$A = \left\langle \prod_{j=1}^{K} f_j^{h_j} \right\rangle = \left\langle \prod_{j=1}^{K} \frac{\nu_j^{h_j}}{\sigma^{h_j}} \right\rangle = \left\langle \frac{1}{\sigma^H} \prod_{j=1}^{K} \nu_j^{h_j} \right\rangle . \quad (60)$$

We make use of the identity

$$\frac{1}{\sigma^H} = \int_0^\infty \frac{x^{H-1}}{(H-1)!} e^{-x\sigma} dx \quad (61)$$

to obtain

$$A = \int_0^\infty \frac{x^{H-1}}{\Gamma(H)} \left\langle e^{-x\sigma} \prod_{j=1}^{K} \nu_j^{h_j} \right\rangle dx. \tag{62}$$

We now use the definition of σ as the sum of the ν_j and separate out the ν_j that correspond to the lineages we are considering. Note that the ν_j are independent of each other. Thus one obtains

$$A = \int_0^\infty \frac{x^{H-1}}{\Gamma(H)} \left\langle e^{-x\sum_{j=K+1}^{B} \nu_j} \right\rangle \left\langle \prod_{j=1}^{K} \nu_j^{h_j} e^{-x\nu_j} \right\rangle dx \tag{63}$$

whence, by independence,

$$A = \int_0^\infty \frac{x^{H-1}}{\Gamma(H)} \left\langle e^{-x\nu_1} \right\rangle^{B-K} \left\langle \prod_{j=1}^{K} \nu_j^{h_j} e^{-x\nu_j} \right\rangle dx. \tag{64}$$

From Eq. (10) we have

$$\left\langle e^{-z\nu_i} \right\rangle = e^{-\mu_i/Uz^{1-1/q}} = e^{-z^\alpha/B}, \tag{65}$$

where $\alpha \equiv 1 - \frac{1}{q}$. Substituting this in, and assuming large B so that $(B-K)/B \approx 1$, we find

$$A = \int_0^\infty \frac{x^{H-1}}{\Gamma(H)} e^{-x^\alpha} \left\langle \prod_{j=1}^{K} \nu_j^{h_j} e^{-x\nu_j} \right\rangle dx. \tag{66}$$

We then use that

$$\left\langle \nu^h e^{-x\nu} \right\rangle = (-1)^h \frac{\partial^h}{\partial z^h} \left[e^{-z^\alpha/B} \right]. \tag{67}$$

Making the large-B approximation that $e^{-z^\alpha/B} \approx 1 - \frac{z^\alpha}{B}$ and differentiating, we find

$$\left\langle \nu^h e^{-x\nu} \right\rangle = \frac{\alpha}{B} \frac{\Gamma(h-\alpha)}{\Gamma(1-\alpha)} x^{\alpha-h}. \tag{68}$$

Using this result, we have

$$A = \int_0^\infty \frac{x^{H-1}}{\Gamma(H)} e^{-x^\alpha} \prod_{j=1}^{K} \frac{\alpha x^{\alpha-h_j} \Gamma(h_j-\alpha)}{B\Gamma(1-\alpha)} dx. \tag{69}$$

Since $\sum_{j=1}^{K} h_j = H$ we can rewrite this as

$$A = \int_0^\infty \frac{x^{K\alpha} dx}{x} \frac{\alpha^K}{B^K \Gamma(H)} e^{-x^\alpha} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(1-\alpha)}. \tag{70}$$

Now we define

$$y = x^\alpha \qquad dy = \alpha x^{\alpha-1} dx \qquad \frac{dy}{\alpha y} = \frac{dx}{x}, \tag{71}$$

and making this change of variables obtain

$$A = \int_0^\infty \frac{dy}{\alpha} \frac{y^{K-1} e^{-y} \alpha^K}{B^K \Gamma(H)} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(1-\alpha)}. \tag{72}$$

The dy integral yields a Γ function, giving

$$A = \frac{\Gamma(K)\alpha^{K-1}}{B^K \Gamma(H)} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(1-\alpha)}. \tag{73}$$

So far we have considered the probability of this coalescence event involving K lineages at a specific set of K sites on the genome. We now want to sum over all the possible sets of K sites on the genome at which this could occur. There are a total of $B^K/K!$ of these. We define E to be the probability of this coalescence event involving K lineages at *any* set of K sites on the genome. We have

$$E = \frac{\alpha^{K-1}}{K\Gamma(H)} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(1-\alpha)}. \tag{74}$$

Now so far we have assumed that specific individuals coalesce into specific lineages. But given a set $\{h_j\}$ there are a total of $\binom{H}{h_1,h_2,\ldots h_K}$ ways to assign specific individuals to specific lineages. Thus the total probability of H individuals coalescing into K lineages, in a specific configuration $\{h_j\}$, which we will call $C_{H,K,\{h_j\}}$, is

$$\begin{aligned} C_{H,K,\{h_j\}} &= \frac{H!}{\prod_{j=1}^{K} h_j!} \frac{\alpha^{K-1}}{K\Gamma(H)} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(1-\alpha)} \\ &= \frac{H\alpha^{K-1}}{K} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(h_j+1)\Gamma(1-\alpha)}, \end{aligned} \tag{75}$$

equivalent to Eq. (15) in the main text.

To compute D_{HK}, we first make the definition

$$f(H,K) = \sum_{\{h_j\}} \prod_{j=1}^{K} \frac{\Gamma(h_j-\alpha)}{\Gamma(h_j+1)\Gamma(1-\alpha)}, \tag{76}$$

and note that

$$D_{HK} = \frac{H}{K\alpha} f(H,K). \tag{77}$$

There is no simple analytic expression for $f(H,K)$. However, we can define its generating function

$$R_f(z) \equiv \sum_{H=0}^{\infty} f(H,K) z^H. \tag{78}$$

Note we are summing from $H = 0$: even though for $H < K$ this is not biologically relevant, it will be useful formally. Now we have

$$R_f(z) = \sum_{H=0}^{\infty} \sum_{\{h_j\}}^{constrained} f(H,K) z^H \tag{79}$$

$$= \sum_{h_1=0}^{\infty} \sum_{h_2=0}^{\infty} \cdots \sum_{h_K=0}^{\infty} f(H,K) z^H.$$

Substituting in for $f(H,K)$, we find

$$R_f(z) = \left[\sum_{h=0}^{\infty} \frac{\alpha \Gamma(h-\alpha) z^h}{\Gamma(h+1)\Gamma(1-\alpha)} \right]^K, \qquad (80)$$

where we have used the fact that the sums over the different h are now independent. Recognizing the Taylor series, we have

$$R_f(z) = \left[1 - (1-z)^\alpha\right]^K, \qquad (81)$$

as quoted in the main text. Note we can also plug in $K = 1$ to recover the result for D_{H1} quoted in Eq. (14).

[1] J. Maynard-Smith and J. Haigh, Genetical Research **23**, 23 (1974).
[2] P. C. Sabeti, S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander, Science **312**, 1614 (2006).
[3] J. M. Akey, Genome Research **19**, 711 (2009).
[4] R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark, Nature Reviews Genetics **8**, 857 (2007).
[5] J. Novembre and A. Di Rienzo, Nature Reviews Genetics **10**, 745 (2009).
[6] J. K. Pritchard, J. K. Pickrell, and G. Coop, Current Biology **20**, R208 (2010).
[7] P. Gerrish and R. Lenski, Genetica **102/103**, 127 (1998), read.
[8] W. Hill and A. Robertson, Genetical Research **8**, 269 (1966).
[9] D. Ridgway, H. Levine, and D. Kessler, J Stat Phys **90**, 191 (1998), read.
[10] I. Rouzine, J. Wakeley, and J. Coffin, PNAS **100**, 587 (2003).
[11] M. M. Desai and D. S. Fisher, Genetics **176**, 1759 (2007).
[12] O. Hallatschek, PNAS **108**, 1783 (2011).
[13] B. H. Good, I. M. Rouzine, D. J. Balick, O. Hallatschek, and M. M. Desai, Proceedings of the National Academy of Sciences **109**, 4950 (2012).
[14] J. de Visser, C. W. Zeyl, P. J. Gerrish, J. L. Blanchard, and R. E. Lenski, Science **283**, 404 (1999), read.
[15] R. Miralles, P. J. Gerrish, A. Moya, and S. F. Elena, Science **285**, 1745 (1999).
[16] J. P. Bollback and J. P. Huelsenbeck, Mol Biol Evol **24**, 1397 (2007).
[17] M. M. Desai, D. S. Fisher, and A. W. Murray, Current Biology **17**, 385 (2007).
[18] K. C. Kao and G. Sherlock, Nature Genetics **40**, 1499 (2008).
[19] S. Park, D. Simon, and J. Krug, J. Stat. Phys. **138**, 381 (2010).
[20] P. D. Sniegowski and P. J. Gerrish, Philosophical Transactions of the Royal Society B: Biological Sciences **365**, 1255 (2010).
[21] J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim, Nature **461**, 1243 (2009).
[22] J. Barrick and R. Lenski, Cold Spring Harbor Symposia on Quantitative Biology **74**, 119 (2009).
[23] D. Gresham, M. M. Desai, C. M. Tucker, H. T. Jenq, D. A. Pai, A. Ward, C. G. DeSevo, D. Botstein, and M. J. Dunham, PLoS Genet **4**, e1000303 (2008).
[24] P. Ralph and G. Coop, Genetics **186**, 647 (2010).
[25] H. Innan and Y. Kim, PNAS **101**, 10667 (2004).
[26] H. A. Orr and A. J. Betancourt, Genetics **157**, 875 (2001).
[27] G. Sella, D. A. Petrov, M. Przeworski, and P. Andolfatto, PLoS Genetics **5**, e1000495 (2009).
[28] G. Coop, J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R. M. Myers, L. L. Cavalli-Sforza, M. W. Feldman, and J. K. Pritchard, PLoS Genetics **5**, 1000500 (2009).
[29] R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. AUTON, G. McVean, . G. Project, G. Sella, and M. Przeworski, Science **331**, 920 (2011).
[30] J. Hermisson and P. S. Pennings, Genetics **169**, 2335 (2005).
[31] P. S. Pennings and J. Hermisson, Molecular Biology and Evolution **23**, 1076 (2006).
[32] P. S. Pennings and J. Hermisson, PLoS Genetics **2**, e186 (2006).
[33] M. Przeworski, G. Coop, and J. Wall, Evolution **59**, 2312 (2005).
[34] J. K. Pritchard and A. Di Rienzo, Nature Reviews Genetics **11**, 665 (2010).
[35] L. M. Chevin and F. Hospital, Genetics **180**, 1645 (2008).
[36] A. M. Hancock, G. Alkorta-Aranburu, D. B. Witonsky, and A. Di Rienzo, Philosophical Transactions of the Royal Society of London B **365**, 2459 (2010).
[37] A. M. Walczak, L. E. Nicolaisen, J. B. Plotkin, and M. M. Desai, Genetics **190**, 753 (2012).
[38] R. Hudson and N. Kaplan, in *Non-neutral evolution: Theories and molecular data*, edited by B. Golding (Chapman and Hall, New York, 1994) pp. 140–153.
[39] I. Gordo, A. Navarro, and B. Charlesworth, Genetics **161**, 835 (2002).
[40] J. Seger, W. A. Smith, J. J. Perry, J. Hunn, Z. A. Kaliszewska, L. L. Sala, L. Pozzi, V. J. Rowntree, and F. R. Adler, Genetics **184**, 529 (2010).
[41] R. A. Neher, B. I. Shraiman, and D. S. Fisher, Genetics **184**, 467 (2010).
[42] C. A. Fogle, J. L. Nagle, and M. M. Desai, Genetics **180**, 2163 (2008).
[43] I. Rouzine, E. Brunet, and C. Wilke, Theoretical Population Biology **73**, 24 (2008).
[44] E. Brunet, I. Rouzine, and C. Wilke, Genetics **179**, 603 (2008).
[45] D. S. Fisher, **submitted** (2012).
[46] M. M. Desai, L. E. Nicolaisen, A. M. Walczak, and J. B. Plotkin, Theoretical Population Biology **81**, 144 (2012).
[47] E. Bolthausen and A. S. Sznitman, Communications in Mathematical Physics **197**, 247 (1998).

[48] B. Derrida and E. Brunet, arXiv **q-bio.PE**, 1202.5997 (2012).
[49] F. Tajima, Genetics **123**, 585 (1989).
[50] W. J. Ewens, *Mathematical Population Genetics: I. Theoretical Introduction* (Springer, New York, NY, 2004).
[51] J. Pitman, Annals of Probability **27**, 1870 (1999).
[52] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier, Europhysics Letters **76**, 1 (2006).
[53] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier, Physical Review E **76**, 041104 (2007).
[54] E. Brunet, B. Derrida, and D. Simon, Physical Review E **78**, 061102 (2008).
[55] J. Berestycki, N. Berestycki, and J. Schweinsberg, Annals of Probability **in press** (2012).
[56] E. Brunet and B. Derrida, Philosophical Magazine **92**, 255 (2012).
[57] S. Goyal, D. J. Balick, E. R. Jerison, R. A. Neher, B. I. Shraiman, and M. M. Desai, Genetics **191**, 1309 (2012).